IJACSA
WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 11 Issue 1

January 2020

ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

SAI

www.ijacsa.thesai.org

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!

Managing Editor
IJACSA
Volume 11 Issue 1 January 2020
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Paper 1: High Performance Computing in Resource Poor Settings: An Approach based on Volunteer Computing
Authors: Adamou Hamza, Azanzi Jiomekong
PAGE 1 – 10

Paper 2: Search Space of Adversarial Perturbations against Image Filters
Authors: Dang Duy Thang, Toshihiro Matsui
PAGE 11 – 19

Paper 3: Towards Security Effectiveness Evaluation for Cloud Services Selection following a Risk-Driven Approach
Authors: Sarah Maroc, Jian Biao Zhang
PAGE 20 – 31

Paper 4: Hybrid Machine Learning Algorithms for Predicting Academic Performance
Authors: Phauk Sokkhey, Takeo Okazaki
PAGE 32 – 41

Paper 5: Proposal of a Sustainable Agile Model for Software Development
Authors: Oscar Antonio Alvarez Galán, José Luis Cendejas Valdéz, Heberto Ferreira Medina, Gustavo A. Vanegas Contreras, Jesús Leonardo Soto Sumuano
PAGE 42 – 49

Paper 6: Pathological Worrying and Artificial Neural Networks
Authors: Carlos Pelta
PAGE 50 – 54

Paper 7: Performance Comparison of CRUD Methods using NET Object Relational Mappers: A Case Study
Authors: Doina Zmaranda, Lucian-Laurentiu Pop-Fele, Cornelia Győrödi, Robert Győrödi, George Pecherle
PAGE 55 – 65

Paper 8: Using the Convolution Neural Network Attempts to Match Japanese Women's Kimono and Obi
Authors: Kumiko Komizo, Noriaki Kuwahara
PAGE 66 – 71

Paper 9: An Enhanced K-Nearest Neighbor Predictive Model through Metaheuristic Optimization
Authors: Allemar Jhone P. Delima
PAGE 72 – 80

Paper 10: Hybrid Algorithm Naive Bayesian Classifier and Random Key Cuckoo Search for Virtual Machine Consolidation Problem
Authors: Yasser Moaly, Basheer A.Youssef
PAGE 81 – 90

Paper 11: Adjacency Effects of Layered Clouds by Means of Monte Carlo Ray Tracing
Authors: Kohei Arai
PAGE 91 – 98

Paper 12: An Intelligent and Adaptive Model for Change Management
Authors: Ali M. Alshahrani
PAGE 99 – 104

Paper 13: A Systematic Review on Students' Engagement in Classroom: Indicators, Challenges and Computational Techniques
Authors: Latha Subramainan, Moamin A. Mahmoud
PAGE 105 – 115

Paper 14: The Effectiveness of Stemming in the Stylometric Authorship Attribution in Arabic
Authors: Abdulfattah Omar, Wafya Ibrahim Hamouda
PAGE 116 – 121

Paper 15: Edge IoT Networking
Authors: Majed Mohaia Alhaisoni
PAGE 122 – 127

Paper 16: Awareness of Ethical Issues when using an e-Learning System
Authors: Talib Ahmad Almseidein, Omar Musa Klaif.Mahasneh
PAGE 128 – 131

Paper 17: A Blockchain based Mobile Money Interoperability Scheme
Authors: Fickson Mvula, Jackson Phiri, Simon Tembo
PAGE 132 – 140

Paper 18: Sign Language Semantic Translation System using Ontology and Deep Learning
Authors: Eman K. Elsayed, Doaa R. Fathy
PAGE 141 – 147

Paper 19: Malicious URL Detection based on Machine Learning
Authors: Cho Do Xuan, Hoa Dinh Nguyen, Tisenko Victor Nikolaevich
PAGE 148 – 153

Paper 20: Noise Reduction in Spatial Data using Machine Learning Methods for Road Condition Data
Authors: Dara Anitha Kumari, A. Govardhan
PAGE 154 – 163

Paper 21: Blockchain-based Electronic Voting System with Special Ballot and Block Structures that Complies with Indonesian Principle of Voting
Authors: Gottfried Christophorus Prasetyadi, Achmad Benny Mutiara, Rina Refianti
PAGE 164 – 170

Paper 22: Mobile Cloud Learning based on User Acceptance using DeLone and McLean Model for Higher Education
Authors: Kristiawan Nugroho, Sugeng Murdowo, Farid Ahmadi, Tri Suminar
PAGE 171 – 178

Paper 23: Open Challenges for Crowd Density Estimation
Authors: Shaya A. Alshaya
PAGE 179 – 187

Paper 24: Model for Measuring benefit of Government IT Investment using Fuzzy AHP
Authors: Prih Haryanta, Azhari Azhari, Khabib Mustofa
PAGE 188 – 192

Paper 25: Link Breakage Time Prediction Algorithm for Efficient Power and Routing in Unmanned Aerial Vehicle Communication Networks
Authors: Haque Nawaz, Husnain Mansoor Ali
PAGE 193 – 199

Paper 26: IoT System for Sleep Quality Monitoring using Ballistocardiography Sensor
Authors: Nico Surantha, C'zuko Adiwiputra, Oei Kurniawan Utomo, Sani Muhamad Isa, Benfano Soewito
PAGE 200 – 205

Paper 27: A Critical Review on Adverse Effects of Concept Drift over Machine Learning Classification Models
Authors: Syed Muslim Jameel, Manzoor Ahmed Hashmani, Hitham Alhussain, Mobashar Rehman, Arif Budiman
PAGE 206 – 211

Paper 28: Scalability Performance for Low Power Wide Area Network Technology using Multiple Gateways
Authors: N.A. Abdul Latiff, I.S. Ismail, M. H. Yusoff
PAGE 212 – 218

Paper 29: Tuberculosis Prevention Model in Developing Countries based on Geospatial, Cloud and Web Technologies
Authors: Innocent Mwila, Jackson Phiri
PAGE 219 – 232

Paper 30: Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List
Authors: Gayatri Venugopal-Wairagade, Jatinderkumar R. Saini, Dhanya Pramod
PAGE 233 – 239

Paper 31: An IoT based Home Automation Integrated Approach: Impact on Society in Sustainable Development Perspective
Authors: Yasir Mahmood, Nazri Kama, Azri Azmi, Suraya Ya'acob
PAGE 240 – 250

Paper 32: Crime Mapping Model based on Cloud and Spatial Data: A Case Study of Zambia Police Service
Authors: Jonathan Phiri, Jackson Phiri, Charles S. Lubobya
PAGE 251 – 265

Paper 33: Classification Models for Determining Types of Academic Risk and Predicting Dropout in University Students
Authors: Norka Bedregal-Alpaca, Víctor Cornejo-Aparicio, Joshua Zárate-Valderrama, Pedro Yanque-Churo
PAGE 266 – 272

Paper 34: Predicting the Future Transaction from Large and Imbalanced Banking Dataset
Authors: Sadaf Ilyas, Sultan Zia, Umair Muneer Butt, Sukumar Letchmunan, Zaib un Nisa
PAGE 273 – 286

Paper 35: SBAG: A Hybrid Deep Learning Model for Large Scale Traffic Speed Prediction
Authors: Adnan Riaz, Muhammad Nabeel, Mehak Khan, Huma Jamil
PAGE 287 – 291

Paper 36: Categorizing Attributes in Identifying Learning Style using Rough Set Theory
Authors: Dadang Syarif Sihabudin Sahid, Riswan Efendi, Emansa Hasri Putra, Muhammad Wahyudi
PAGE 292 – 298

Paper 37: Facial Emotion Recognition using Neighborhood Features
Authors: Abdulaziz Salamah Aljoloud, Habib Ullah, Adwan Alownie Alanazi
PAGE 299 – 306

Paper 38: A New Solution to Protect Encryption Keys when Encrypting Database at the Application Level
Authors: Karim El bouchti, Soumia Ziti, Fouzia Omary, Nassim Kharmoum
PAGE 307 – 312

Paper 39: Neural Network Supported Chemiresistor Array System for Detection of NO2 Gas Pollution in Smart Cities (NN-CAS)
Authors: Mahmoud Zaki Iskandarani
PAGE 313 – 318

Paper 40: A Framework for Detecting Botnet Command and Control Communication over an Encrypted Channel
Authors: Zahian Ismail, Aman Jantan, Mohd. Najwadi Yusoff
PAGE 319 – 326

Paper 41: The Multi-Class Classification for the First Six Surats of the Holy Quran
Authors: Nouh Sabri Elmitwally, Ahmed Alsayat
PAGE 327 – 332

Paper 42: Cancelable Face Template Protection using Transform Features for Cyberworld Security
Authors: Firdous Kausar
PAGE 333 – 341

Paper 43: Development of Flipbook using Web Learning to Improve Logical Thinking Ability in Logic Gate
Authors: Rizki Noor Prasetyono, Rito Cipta Sigitta Hariyono
PAGE 342 – 348

Paper 44: Automatic Detection and Correction of Blink Artifacts in Single Channel EEG Signals
Authors: G Bhaskar N Rao, Anumala Vijay Sankar, Peri Pinak Pani, Aneesh Sidhireddy
PAGE 349 – 353

Paper 45: Usability of Mobile Assisted Language Learning App
Authors: Kashif Ishaq, Nor Azan Mat Zin, Fadhilah Rosdi, Adnan Abid, Qasim Ali
PAGE 354 – 363

Paper 46: Classification of Non-Discriminant ERD/ERS Comprising Motor Imagery Electroencephalography Signals
Authors: Zaib unnisa Asi, M. Sultan Zia, Umair Muneer Butt, Aneela Abbas, Sadaf Ilyas
PAGE 364 – 375

Paper 47: Barchan Sand Dunes Collisions Detection in High Resolution Satellite Images based on Image Clustering and Transfer Learning
Authors: M. A. Azzaoui, L. Masmoudi, H. El Belrhiti, I. E. Chaouki
PAGE 376 – 383

Paper 48: Usefulness of Mobile Assisted Language Learning in Primary Education
Authors: Kashif Ishaq, Nor Azan Mat Zin, Fadhilah Rosdi, Adnan Abid, Qasim Ali
PAGE 384 – 395

Paper 49: Detecting Flooding Attacks in Communication Protocol of Industrial Control Systems
Authors: Rajesh L, Penke Satyanarayana
PAGE 396 – 401

Paper 50: An Artificial Deep Neural Network for the Binary Classification of Network Traffic
Authors: Shubair A. Abdullah, Ahmed Al-Ashoor
PAGE 402 – 408

Paper 51: An Improved Framework for Content-based Spamdexing Detection
Authors: Asim Shahzad, Hairulnizam Mahdin, Nazri Mohd Nawi
PAGE 409 – 420

Paper 52: Combining 3D Interpolation, Regression, and Body Features to build 3D Human Data for Garment: An Application to Building 3D Vietnamese Female Data Model
Authors: Tran Thi Minh Kieu, Nguyen Tung Mau, Le Van, Pham The Bao
PAGE 421 – 429

Paper 53: Breast Cancer Computer-Aided Detection System based on Simple Statistical Features and SVM Classification
Authors: Yahia Osman, Umar Alqasemi
PAGE 430 – 433

Paper 54: EEG Emotion Signal of Artificial Neural Network by using Capsule Network
Authors: Usman Ali, Haifang Li, Rong Yao, Qianshan Wang, Waqar Hussain, Syed Badar ud Duja, Muhammad Amjad, Bilal Ahmed
PAGE 434 – 443

Paper 55: Behavior of Learning Rules in Hopfield Neural Network for Odia Script
Authors: Ramesh Chandra Sahoo, Sateesh Kumar Pradhan
PAGE 444 – 452

Paper 56: Critical Analysis of Brain Magnetic Resonance Images Tumor Detection and Classification Techniques
Authors: Zahid Ullah, Su-Hyun Lee, Donghyeok An
PAGE 453 – 465

Paper 57: Secure V2V Communication in IOV using IBE and PKI based Hybrid Approach
Authors: Satya Sandeep Kanumalli, Anuradha Ch, Patanala Sri Rama Chandra Murty
PAGE 466 – 472

Paper 58: A Multi-Objectives Optimization to Develop the Mobile Dimension in a Small Private Online Course (SPOC)
Authors: Naima BELARBI, Abdelwahed NAMIR, Mohamed TALBI, Nadia Chafiq
PAGE 473 – 480

Paper 59: A Fuzzy Multi-Objective Covering-based Security Quantification Model for Mitigating Risk of Web based Medical Image Processing System
Authors: Abdullah Algarni, Masood Ahmad, Abdulaziz Attaallah, Alka Agrawal, Rajeev Kumar, Raees Ahmad Khan
PAGE 481 – 489

Paper 60: Knowledge Sharing Factors for Modern Code Review to Minimize Software Engineering Waste
Authors: Nargis Fatima, Sumaira Nazir, Suriayati Chuprat
PAGE 490 – 497

Paper 61: Situational Factors for Modern Code Review to Support Software Engineers' Sustainability
Authors: Sumaira Nazir, Nargis Fatima, Suriayati Chuprat
PAGE 498 – 504

Paper 62: Plant Disease Detection using Internet of Thing (IoT)
Authors: Muhammad Amir Nawaz, Tehmina khan, Rana Mudassar Rasool, Maryam Kausar, Amir Usman, Tanvir Fatima Naik Bukht, Rizwan Ahmad, Jaleel Ahmad
PAGE 505 – 509

Paper 63: Performance Analysis of Machine Learning Classifiers for Detecting PE Malware
Authors: ABM.Adnan Azmee, Pranto Protim Choudhury, Md. Aosaful Alam, Orko Dutta, Muhammad Iqbal Hossai
PAGE 510 – 517

Paper 64: Determinants of Interface Criteria Learning Technology for Disabled Learner using Analytical Hierarchy Process
Authors: Syazwani Ramli, Hazura Mohamed, Zurina Muda
PAGE 518 – 523

Paper 65: Predicting IoT Service Adoption towards Smart Mobility in Malaysia: SEM-Neural Hybrid Pilot Study
Authors: Waqas Ahmed, Sheikh Muhamad Hizam, Ilham Sentosa, Habiba Akter, Eiad Yafi, Jawad Ali
PAGE 524 – 535

Paper 66: Robotic Technology for Figural Creativity Enhancement: Case Study on Elementary School
Authors: Billy Hendrik, Nazlena Mohamad Ali, Norshita Mat Nayan
PAGE 536 – 543

Paper 67: The Impact of Deep Learning Techniques on SMS Spam Filtering
Authors: Wael Hassan Gomaa
PAGE 544 – 549

Paper 68: Translator System for Peruvian Sign Language Texts through 3D Virtual Assistant
Authors: Gleny Paola Gamarra Ramos, María Elisabeth Farfán Choquehuanca
PAGE 550 – 555

Paper 69: A Service-Oriented Architecture for Optimal Service Selection and Positioning in Extremely Large Crowds
Authors: Mohammad A.R. Abdeen
PAGE 556 – 561

Paper 70: EDUGXQ: User Experience Instrument for Educational Games' Evaluation
Authors: Vanisri Nagalingam, Roslina Ibrahim, Rasimah Che Mohd Yusoff
PAGE 562 – 569

Paper 71: Stemming Text-based Web Page Classification using Machine Learning Algorithms: A Comparison
Authors: Ansari Razali, Salwani Mohd Daud, Nor Azan Mat Zin, Faezehsadat Shahidi
PAGE 570 – 576

Paper 72: Performance Realization of CORDIC based GMSK System with FPGA Prototyping
Authors: Renuka Kajur, K V Prasad
PAGE 577 – 584

Paper 73: HybridFatigue: A Real-time Driver Drowsiness Detection using Hybrid Features and Transfer Learning

Authors: Qaisar Abbas

PAGE 585 – 593

Paper 74: An Improved Deep Learning Approach based on Variant Two-State Gated Recurrent Unit and Word Embeddings for Sentiment Classification

Authors: Muhammad Zulqarnain, Suhaimi Abd Ishak, Rozaida Ghazali, Nazri Mohd Nawi, Muhammad Aamir, Yana Mazwin Mohmad Hassim

PAGE 594 – 603

Paper 75: Angle Adjustment for Vertical and Diagonal Communication in Underwater Sensor

Authors: Aimen Anum, Tariq Ali, Shuja Akbar, Iqra Obaid, Muhammad Junaid Anjum, Umar Draz, Momina Shaheen

PAGE 604 – 616

Paper 76: Towards an Intelligent Approach for the Restitution of Physical Soil Parameters

Authors: Ibtissem HOSNI, Lilia BENNACEUR FARAH, Imed Riadh FARAH, Raouf BENNACEUR, Mohamed Saber NACEUR

PAGE 617 – 624

Paper 77: An Innovative Smartphone-based Solution for Traffic Rule Violation Detection

Authors: Waleed Alasmary

PAGE 625 – 636

Paper 78: HarmonyMoves: A Unified Prediction Approach for Moving Object Future Path

Authors: Mohammed Abdalla, Hoda M. O. Mokhtar, Neveen ElGamal

PAGE 637 – 644

Paper 79: A Survey on Cloud Data Security using Image Steganography

Authors: Afrah Albalawi, Nermin Hamza

PAGE 645 – 649

Paper 80: Developing Decision Tree based Models in Combination with Filter Feature Selection Methods for Direct Marketing

Authors: Ruba Obiedat

PAGE 650 – 659

Paper 81: Modelling an Indoor Crowd Monitoring System based on RSSI-based Distance

Authors: Syifaul Fuada, Trio Adiono, Prasetiyo, Hartian Widhanto Shorful Islam

PAGE 660 – 667

Paper 82: Priority-based Routing Framework for Image Transmission in Visual Sensor Networks: Experimental Analysis

Authors: Emad Felemban, Atif Naseer, Adil Amjad

PAGE 668 – 677

Paper 83: A Deep Learning Approach for Handwritten Arabic Names Recognition

Authors: Mohamed Elhafiz Mustafa, Murtada Khalafallah Elbashir

PAGE 678 – 682

Paper 84: Optimal Topology Generation for Linear Wireless Sensor Networks based on Genetic Algorithm

Authors: Adil A. Sheikh, Emad Felemban

PAGE 683 – 692

Paper 85: DDoS Flooding Attack Mitigation in Software Defined Networks

Authors: Safaa MAHRACH, Abdelkrim HAQIQ

PAGE 693 – 700

Paper 86: Investigation of Deep Learning-based Techniques for Load Disaggregation, Low-Frequency Approach

Authors: Abdolmaged Alkhulaifi, Abdulah J. Aljohani

PAGE 701 – 706

Paper 87: Pedestrian Crowd Detection and Segmentation using Multi-Source Feature Descriptors

Authors: Saleh Basalamah, Sultan Daud Khan

PAGE 707 – 713

Paper 88: Towards an Improvement of Fourier Transform

Authors: Khalid Aznag, Toufik Datsi, Ahmed El Oirrak, Essaid El Bachari

PAGE 714 – 719

Paper 89: LEA-SIoT: Hardware Architecture of Lightweight Encryption Algorithm for Secure IoT on FPGA Platform

Authors: Bharathi R, N. Parvatham

PAGE 720 – 725

Paper 90: Delay-Aware and User-Adaptive Offloading of Computation-Intensive Applications with Per-Task Delay in Mobile Edge Computing Networks

Authors: Tarik Chanyour, Youssef HMIMZ, Mohamed EL GHMARY, Mohammed Oucamah CHERKAOUI MALKI

PAGE 726 – 733

High Performance Computing in Resource Poor Settings: An Approach based on Volunteer Computing

Adamou Hamza[1], Azanzi Jiomekong[2]
University of Yaounde I, Faculty of Sciences, Yaounde, Cameroon;
IRD, Sorbonne Université, UMMISCO, F-93143, Bondy, France

***Abstract*—High Performance Computing (HPC) systems aim to solve complex computing problems (in a short amount of time) that are either too large for standard computers or would take too long. They are used to solve computational problems in many fields such as medical science (for drug discovery, breast cancer detection in images, etc.), climate science, physics, mathematical science, etc. Existing solutions such as HPC Supercomputer, HPC Cluster, HPC Cloud or HPC Grid are not adapted for resource poor settings (mainly for developing countries) because their fees are generally beyond the funding (particularly for academics) and the administrative complexity to access to HPC Grid creates a higher barrier. This paper presents an approach allowing to build a Volunteer Computing system for HPC in resource poor settings. This solution does not require any additional investment in hardware, but relies instead on voluntary machines already owned by the private users. The experiment has been made on the mathematical problem of solving the matrices multiplication using Volunteer Computing system. Given the success of this experiment, the enrollment of other volunteers has already started. The goal being to create a powerful Volunteer Computing system with the maximum number of computers.**

Keywords*—*Volunteer computing; resource poor settings; high performance computing; matrix multiplication

I. Introduction

High performance computing [1], [2], [3], [4], [5] is really important in most scientific and industrial sectors. It helps scientists gain valuable insights to boost innovation and discovery in almost all areas of science ranging from life sciences [6], [7], [8] to quantum mechanic [9] and large scale data mining [10], [11]. In the industrial sector, high performance computing gives companies a significant competitive advantage in reducing the costs of development cycles and in producing higher quality products and services [12]. Societal challenges, including preventing and managing natural disasters, early detection and treatment of disease, and forecasting climate evolution require very high computing power [13].

In the past, dedicated supercomputers [14], [15] powerful machines made up of a large number of processors, were only used to build a HPC system. Despite the processing capacity and the speed of calculation of these computers, they do not offer the commodity price advantage, especially for small businesses and academics [13], [16]. Then, the following solutions were proposed : HPC Clusters [17], [18], [19], HPC Cloud [20], [21], and HPC Grid [19], [22]. HPC Cluster is a system composed of computers connected to a local area network, while HPC Cloud is a system involving computers connected in a private or public network. Finally, HPC Grid is composed of computers connected in wide-area networks such as Internet.

The proposed solutions cannot be applied in low resource areas because of its costs. The HPC Cluster and the HPC Cloud require financial resources beyond the reach of most institutions in developing countries. For the HPC Grid, administrative complexity in obtaining the necessary permissions to use this system can be a higher barrier [21].

In developed countries, HPC have provided computing resources to projects at a huge scale [23], often resulting in major scientific discoveries and invention whether at the universities or businesses level. However, in resource poor or limited settings, many academics and small business do not always have enough resources to access to the HPC. Resource poor or constrained settings are defined as a local where the capability to provide HPC is limited to basic critical resources, including desktops and laptops. Resource poor settings can be stratified by no resources and limited resources. In resource limited settings, people use their personal computers for computation. This article focuses on the limited resources category. These include scientists in research teams with significant computational needs, individuals in developing countries with no access to other alternatives.

It should be noted that in many developing countries, the wide range of commercial centers of computers has made available many devices. For instance, in the Department of Computer Science at the University of Yaounde I in Cameroon, there are about 1,000 computers owned by lecturers and students; the Internet connection is widespread in Cameroon. These computers are idle most of the time (e.g., when students are in class or sleeping at night, etc.), this therefore constitutes a source of computing power available and largely reusable. If their owners are willing to actively lend CPU time and memory, these devices can be used as distributed computing infrastructure at no cost. This is called a Volunteer Computing (VC) system [24], [25], [26], [27]. A HPC Volunteer Computing system is obtained through community engagement by setting up a system of volunteer machines. Their goal is similar to HPC Grid, which is to gather distributed computing resources and federate them to solve large computational problems. The difference between these two systems is that the resources come from non-dedicated computers, underutilized and controlled by their owners (volunteers). This approach requires no additional hardware investment, but relies on devices already owned by users and their communities.

This article presents an approach to build a Volunteer Computing system for HPC in resource poor settings. Initially, the Section II describes the main solutions to HPC problems. Then, the Sections III and IV follow with our solution and the experience of this solution respectively. Finally, the Section V concludes and opens future directions.

II. High Performance Computing (HPC)

High Performance Computing uses the principle of parallel computing to address the high computing requirements of applications by dividing them into smaller ones that can be processed simultaneously on different computing units. These computing units can reside on the same computer (Supercomputer) or on multiple computers connected by a network forming HPC Cluster, HPC Cloud, HPC Grid and HPC Volunteer Computing. These solutions are presented in this order in the rest of this section.

A. HPC Supercomputers

One of the best known type of HPC solution is the Supercomputer [14], [15], [28]. A Supercomputer contains hundreds, thousands or even millions of computing units forming a massively paralleled processor organized in a network of processors [29], [30]. For instance, the *Summit - IBM Power System AC922*[1] contains $2,414,592$ cores. These processors work together to solve large computational problems as efficiently and quickly as possible.

Supercomputers allow us to obtain a great computing power. However, with the entry fees [13], [15], acquiring a Supercomputer is almost impossible for scientists and engineers working in resource poor settings. Because of this limitation, other HPC solutions that do not require a fully dedicated computer have been developed. These solutions are HPC Cluster, Cloud, Grid and Volunteer Computing.

B. HPC Cluster

A HPC cluster is a computing system in which independent computers are connected by a high performance local area network in order to solve complex problems [18], [19], [31], [32]. Each machine in the HPC cluster is a complete computer consisting of one or more CPUs or cores, memory, disk drives and network interfaces. HPC cluster is generally owned by a single administrative entity. The software used to manage clusters give users the illusion that they are with a single large computer when in reality the cluster may consists of hundreds or thousands of individual machines [33]. A cluster is much more cost-effective than a single supercomputer of comparable speed [13], [16].

An example of a computer cluster is the dedicated physical cluster at HP Labs Singapore (HPLS)[2]. This cluster is connected to a Gigabit Ethernet network on a single switch. Each server, with 2 CPU sockets (populated with a 6 core CPU), results in twelve physical cores per machine.

[1] https://www.top500.org/system/179397
[2] https://xrds.acm.org/article.cfm?aid=2000789

C. HPC Cloud

A High Performance Computing Cloud [20], [21], [34], [35] is an on-demand availability of computing power, without direct active management by the user. It provides dynamic and scalable computing power through resources organized in data centers distributed around the world. By purchasing on-the-go and not as an asset, HPC Cloud delivers consistent, scalable results, minimizing the initial costs of computing infrastructure. A cloud can be private (operating for a single organization) or public (open for public use). Since 2016, many IT companies such as Amazon Web Services (AWS)[3], Microsoft Azure cloud[4], Google Cloud Platform[5] have offered HPC Cloud.

The advantage of the HPC cloud over other types of HPC is that it allows the user to immediately access computing resources without the approval of an allocation committee and the service can be provided without human interaction with the service provider. Software can be used without the need to purchase a license or install it, and users do not need to have strong software/infrastructure management skills [21], [36]. However, in the context of resource poor settings, the main drawback of HPC Cloud is that it relies on dedicated hardware managed centrally, which implies a minimum cost which can be high for academics and small businesses [21].

D. HPC Grid

A Grid consists of many computing resources (from multiple administrative domains) connected to a network (e.g., The Internet) working together to solve large problems requiring HPC. The whole system is called a HPC Grid [19], [22], [37]. The HPC Grid differs from the HPC Cluster or Cloud in that it uses many computers, but with a much more distributed nature. Some HPC Grids span the world while others are located within a single organization.

The HPC Grid capabilities are generally managed by a precise organization and computational resources are provided by various supporting institutions, such as companies, research groups, laboratories, and universities. Large Grid computer facilities are often used by a large number of users to solve intensive scientific, mathematical, and academic problems. However, the administrative complexity (obtaining necessary authorizations) allowing the public to access a HPC Grid is a real barrier for academics and small businesses. [25].

An example of Grid is the French Grid'5000[6]. This Grid is distributed over 8 sites, 31 clusters and 12,328 cores. Each site hosts a cluster and all sites are connected by high speed network. It aims to provide a highly reconfigurable, controllable and monitorable experimental platform for research in large-scale parallel and distributed systems.

E. HPC Volunteer Computing

In HPC Volunteer Computing (VC) [24], [25], [27], [38], [39], computer users/owners, who are members of the general public contribute their computing resources to solve HPC

[3] https://aws.amazon.com
[4] https://azure.microsoft.com
[5] https://cloud.google.com
[6] https://www.grid5000.fr/w/Grid5000:Home

problems. It is based on two pillars: the first is the allocation and management of large computing tasks; the second is the participation of a large number of individuals volunteer who offer their computing resources to a project. Compared to HPC Cluster, HPC Cloud and HPC Grid, HPC VC removes financial and administrative barriers. The costs are supported by volunteers, which cover the acquisition, operation, and maintenance of the computing devices.

HPC Volunteer Computing has produced many remarkable scientific results over the last decade. The most popular Volunteer Computing systems are: SETI@home[7][8] [40] for searching the extraterrestrial intelligent life and Folding@home[9] [41] for statistical calculations of molecular dynamics trajectories for models of biological systems. The Folding@home project is a good example of how important scientific results can be produced with VC for affordable problems for other HPC schemes. For instance, in 2008, it was used to study mutations of influenza hemagglutinin [23].

The main challenges of the Volunteer Computing approach is the capabilities of personal devices, the need to encourage and maintain volunteer engagement, and the automatic management of volunteer unreliability [42], [43]. Despite the previous challenges, Volunteer Computing is the solution for HPC in resource poor settings. In fact, in many developing countries, the wide range of commercial centers of computers has made available a lot of devices. These devices spend a lot of time without being used. They can then be used free of charge as a distributed computing infrastructure. It requires no investment in additional hardware, rather relies on devices that generally belong to users and their community, and favours simple tools that can be implemented part-time by a single developer. Section III, will be concerned with a methodology for deploying and using HPC to improve computing in resource poor settings while in Section IV, it will be shown in experiments that this solution can solve the problem of HPC in resource poor settings.

III. An approach based on Volunteer Computing for HPC in resource poor settings

The lack of HPC resources is a challenge for scientists, engineers and businesses in resource poor settings. However many countries have an Internet or local network and many computers are owned by individuals. For example, in the computer science department of the University of Yaounde I in Cameroon, there are around 1,000 computers belonging to students and lecturers. Generally, these computers are not used full time. For example, students spend a lot of time a week attending classes, sleeping, or taking breaks. These computers can be used during their idle time to build a platform for Volunteer Computing. Considering the resource poor settings constraints, this section presents an approach (summarized in Fig. 1) for HPC in these environments. This approach is divided into three main activities: management activity (Section III-A), processing activity (Section III-B) and support activity (Section III-C).

[7]https://setiathome.berkeley.edu/
[8]https://www.seti.org/
[9]https://foldingathome.org/

A. The Management Activity

The management activity uses information about the daily work of volunteers (e.g., the performance of the system and of each volunteer) to supervise the system and increase efficiency. This is the essential key to the success of the Volunteer Computing system. For instance, information about the volunteers, the types of tasks performed (e.g. matrix multiplication, polynomial multiplication), the name of each task, the duration of execution of the tasks, the success or failure of the execution of the tasks and the dates of their execution, will allow us to consider the resource poor settings context. The management activity involves the following activities: design of the Volunteer Computing system (Section III-A1), planning (Section III-A2), coordination (Section III-A3), staffing (Section III-A4), motivation (Section III-A5), control (Section III-A6) and quality assurance (Section III-A7).

1) Design of the Volunteer Computing System: The first and main activity of this approach is to set up the Volunteer Computing system. Resource poor settings generally suffer from power outages and malfunctions in Internet connectivity. In our approach, potential volunteers are firstly users who are often connected to the Internet, secondly those who need the system later and finally friends, family members or communities. Each volunteer has information about the others. A server is designed and contains the most up-to-date information about the volunteers, sent by the latter. All other local information for volunteers is updated by the server. If the server crashes (due to a power outage or a failure), the election algorithm [44], [45] is used to choose the volunteer that will replace the server by waiting that the problem is solved. In our case, it will consist of choosing the volunteer who connects the most to the Internet. Overall, the proposed VC is defined by the equation 1, where:

- S is the server. In our approach, the server is used to centralize all the volunteer information in the system in order to restore it whenever a volunteer needs it. Thus, the server will be used to store all the information on the system, the performance of the system, and of each volunteer in the system, the logs on the computation performed, the type of tasks already completed and its performances, tasks in progress or waiting. These information are essential for efficient management of the system. The server will use the feedback to help volunteers estimate the execution time each time a volunteer wants to perform a task by the system.
- V_i a volunteer computer. It receives tasks, performs them and returns the result. It can submit tasks to other volunteers, retrieve the results and merge to obtain the final result. It informs the server of all the computation activities carried out, in progress or pending. It receives updated information from other volunteers from the server.
- P_i the profile of the volunteer V_i. The profile P_i contains information on the elements that can be used to determine its computing capacity. These are: CPU speed (cpu_speed), number of cores (num_core), memory size (mem_size), disk size (disk_size), operating system name (os_name), its location and its

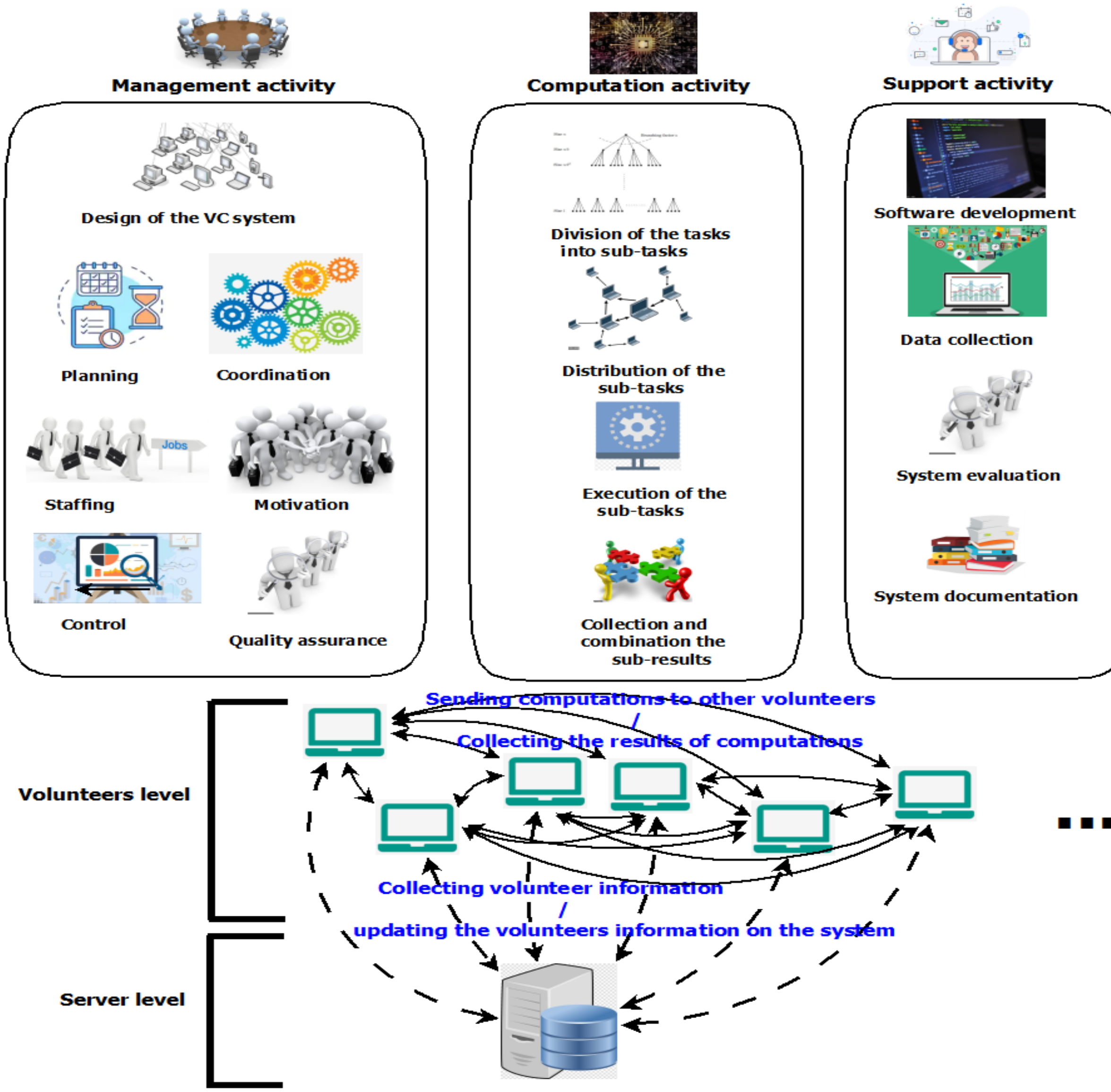


Fig. 1. The Volunteer Computing approach

availability (avail). These information will be used by each volunteer to identify to which volunteers they can send their computations to, what types of computations to send and when.

$$\begin{aligned} VC &= [S, (V_1, P_1), ..., (V_i, P_i)], \\ P_i &= [cpu\_speed, num\_core, mem\_size, \\ &\quad disk\_size, os\_name, location, avail]. \end{aligned} \quad (1)$$

A software named VCSoftware is used in the system. It is composed of the volunteersManager module and the serverManager module:

- volunteersManager module: The volunteerManager module is used at the volunteer level. A new volunteer signs up to the system by filing out their profile in a form, installing the VCSoftware and activating the volunteerManager on their machine. The volunteers-Manager is used to collect information about each volunteer and send it to the server, receives tasks from other volunteers, performs and returns the result, allocate tasks to other volunteers, collect and merge the results and send feedback of computation to the server. The volunteer contribution level is determined by the setting of their profile. The participant can choose to contribute permanently or only when the computer is idle, or can decide whether to contribute or not when the computer is running on battery/inverter. If the server is down, this software informs the computer owner and all other volunteers, the goal being to design the new server.
- serverManager module: The serverManager module is used at the server level. It allows to periodically collect information on the profile of each volunteer and to record a log of all the tasks performed by the system.

This information will allow to recommend a profile to a volunteer (e.g. computation time taking into account the idle period of the computer), to know the performance of each volunteer and the performance of the whole system. When the server is down, the volunteer elected as server while waiting for the problem to be solved activates this module.

2) Planning: Planning activity is the foundation of management. It refers of determining the future course of action towards the desired objective of the system. The planning activity must anticipate and precede all the other functions of the management activity and permit to meet the challenges of environmental changes (e.g., arrival and departure of volunteers, server failure, power/Internet outage etc.). In our VC approach the planning activity is at two levels:

- At the server level, planning involves continuous assessment of the strengths and weaknesses of the VC system using information about volunteers and the results of computation; identification of the actions assigned to a task already performed by the system. It requires all the information collected on the volunteers during the support activity (see Section III-C). This information will allow the server to know the idle time of each volunteer computer, when each one connects to the Internet, to evaluate the performance of each volunteer and the performance of the whole system. For example, information about a task already executed by the system can be used to efficiently execute it the next time the same type of task is submitted.
- At the volunteer level, planning requires the information from other volunteers. Stored on the server, this information is sent to each volunteers when they connect to the system. This information will be used to decide which volunteer to send a task taking into account the computing power offered by each volunteer and the period of connection to the system: which task to send, when to send the task, to whom the task can be sent, the duration of each task sent to a volunteer and the duration of the whole task. At the volunteer level, good planning will help to effectively address the challenges of environmental change (e.g., when a volunteer's computer cannot complete a sub-task).

3) Coordination: Like the planning activity, the coordination activity is done at two levels:

- At the server level, the server identifies the volunteers available for the computation, the computation power they offer and the duration for which they remain connected to the Internet. It also stores information about the different types of tasks already performed. This information will be used to predict the time and resources required to complete a task. Any change in volunteer information (computing power) should result in organizational changes. Information about leaving/joining the system by a volunteer must be considered.
- At the volunteer level, the coordination activity involves organizing the sub-tasks to be performed, as well as the time and resources necessary to carry them out. It will use the information obtained from the other volunteers to identify the resources needed to achieve a given task.

4) Staffing: The staffing activity includes recruiting good volunteers, selecting a group of volunteers to perform a task, and evaluating the volunteers registered in the system. Since volunteers registered in the system are not paid, the manager must be careful during their recruitment. In a resource poor settings, we recommend starting with people who need High Performance Computing and who do not have enough financial resources; and encourage them to invite members of their community to register. Information on volunteer profiles will be used to identify the good volunteers profiles.

5) Motivation: The motivation activity consists of attracting volunteers to contribute to the system and those in the system to increase their contributions. Since volunteers are not paid, a motivational environment must be created. In our case, the possibility of having access to a High Performance Computing is a great motivation for students and researchers. Data collected on the use of the system by other volunteers must be made available to the public, mainly students, engineers and researcher in order to encourage them to join the system. Performance data desired by the system will be made available to volunteers enrolled in the system to encourage them to participate and to encourage members of their communities to participate.

6) Control: The control activity aims to guarantee that scheduled tasks are completed as planned. During the control activity, the performance evaluation of each volunteer and the whole system is made in order to identify weaknesses and strengths. Planning is the basis of control. It focuses on the tasks that are performed and the results of those tasks. It plays an important role in ensuring the efficiency and effectiveness of the VC system. The information collected during the use of the VC system will help to measure the performance of each volunteer but also to identify gaps. At the server level, the comparison between planned performance and actual performance is analyzed to know if there are deviations, and the reasons of the deviations are analyzed. At the volunteer level, each result of a sub-task collected from volunteers is used to evaluate it. Real time information will allow quick control, which will reduce the costs of planning errors.

7) Quality assurance: The quality assurance activity guarantees that the quality of each computation is satisfactory (took the expected time, returned the expected results).

B. Computation Activity

To perform a task with the VC, that task must be divided into sub-tasks and each sub-task sent to volunteers. After the execution by the volunteers, the results and the results logs are collected by the volunteer who initiated the computation. Globally, computation activity involves: dividing the task into sub-tasks, distributing of sub-tasks to volunteers, performing of sub-tasks by volunteers, collecting and combining the results.

1) Division of the task into sub-tasks: This step consists of dividing the task into smaller and independent sub-tasks (preferably atomic sub-tasks). Consider T as a given task to be performed by a volunteer. Then, $T = t_1, t_2, ..., t_n$ where n is the number of independent tasks that compose the task T. Since t_i are independent, they can be executed in parallel.

2) Distribution of the sub-tasks amongst volunteers: Given the data obtained from the server on the system (computing power offered by each volunteer, time to connect to the Internet of each volunteer, etc.), the volunteersManager will identify the volunteers that can participate in the computation. These are the computers most likely to be available until the end of a given sub-task. Subsequently, the computation will be sent to these volunteers when they connect to the system. Depending on the number of volunteers available and the computation time they provide, many sub-tasks can be sent to a volunteer.

The volunteer computer that sent the tasks to others will send a log file to the server to inform it that a job has started. The server will broadcast this information to all other volunteers. If a machine starts a new task, it must consider the existing tasks running in the system.

3) Execution of the sub-tasks: The volunteersManager on the volunteer machine that receives a task will start within the period specified in the volunteer profile. If many tasks have been assigned to a volunteer, the volunteersManager will perform the oldest first. At the end, a log records information on the execution time and the execution status (finished or failed).

4) Collection of the sub-results: After completion, the volunteersManager collects the results and the result logs. If some computations failed, more efficient volunteers are identified and these computations are sent to them. At the end of the computation, the results logs are sent to the server.

5) Combination of the sub-results: The last activity of the computation activity is the combination of the sub-results obtained from the volunteers.

C. Support Activity

The support activity involves the series of activities performed at the same time as the management and the computation activities. It aims to facilitate management and computation by providing all the needed tools and information. The VCSoftware software is developed during this activity. Overall, this activity includes the software development (section III-C1), data collection on the system and each volunteer computer (section III-C2), system evaluation (section III-C3) and system documentation (section III-C4).

1) Software development: The software development activity involves the development/updating of the VCSoftware that will be used by the server and the volunteers. This tool is composed of two main modules: the serverManager module and the volunteersManager module.

The serverManager will help acquire data sent by volunteers to the server. These data are those provided by the volunteer during registration, but also other information on the idle time of the volunteer computer and the time of connection to the Internet. The Information collected from the volunteers will allow the server to: make suggestions of profiles to the volunteers. The volunteer can use the suggestion or not. For example, during the holidays, the idle period is not necessarily the same as during the working period. Then, during the holidays, the server can suggest to volunteers to update their profile.

The volunteersManager is the module which, on the volunteer side, will: periodically collect information on the volunteer and send it to the server; allow identification of volunteer computers that can perform a given task, send tasks to volunteers and collect the results; reception of tasks and their scheduling according to other tasks already received; and sending logs on the execution of the tasks and the performance of each volunteers.

2) Data collection: The data collection activity is done both at the server and at the volunteer level. At the server level, the data collected is used to evaluate the performance of the system and each volunteer. The performance of the whole system is calculated based on the number of volunteers registered in the system, the tasks performed by the volunteers, and when the volunteers will connect to the Internet.

At the volunteer computer level, data is collected in two cases: firstly, the volunteer fills out a registration form in which information about the device and its availability for computation is provided. Then, an automatic data collection takes place (using volunteersManager module). This can be done periodically (hourly or daily) and will concern the idle time of the computer and the time when the computer is connected to the Internet. This data is used to: suggest profile updates to the volunteer and evaluate each volunteer. Each time the server receives updated information from a volunteer, this information is aggregated with existing information and forwarded to other volunteers.

3) System evaluation: The system evaluation involves the evaluation of the server and each volunteers. Indeed data collected during the computations will allow the server to know if the system is efficient and if it provides relevant computations. For this purpose, during the execution of each task, the name, the type, the execution time of the task, the date of execution and the status (execution failed or not) are recorded in a log file and sent to the server.

4) System documentation: The system documentation activity is the vital and continuous activity of our approach. In fact, during this activity, the documents are produced to help the manager and the volunteers to use the system. It provides all information on the capabilities and characteristics of the system. It helps users understand the system and its essential reference materials. VC documentation is updated every time there are changes in the system. The documentation activity is at two levels: the manager level and the volunteer level.

At the manager level, the documents will allow the manager to install and configure the server. It will also give tips to the manager on how to motivate the volunteers to register, contribute and maximize system performance as well as how to customize the software so that it works best for each volunteers (e.g., profile suggestion given data collected from volunteers). The manager can also write Frequently Asked Questions (FAQ) to help volunteers.

At the volunteer level, documents are used to inform the volunteer about the system, and describe what it is intended to do and how it works. Overall, it explains to volunteers how to install the VCSoftware in order to contribute/use the system and how to submit computations and collect the results. To facilitate access to non IT volunteers, a short video is a good way to show how to install the VCSoftware and another on how to submit tasks and collect results. The documentation about the contribution of other users is provided to newcomers in order to motivate the latter.

IV. Experimentation

For Volunteer Computing to be adopted, volunteer devices must provide enough computing power to solve associated computing problems. The first step in the development of our VC system consists of a pilot phase, which is the development and experimentation of a VC system. This section shows how this system has been built and used to solve the problem of matrix multiplication. In the following, the problem of matrix multiplication is first described in section IV-A. Section IV-B follows with the VC_UY Volunteer Computing system built at the University of Yaounde I in Cameroon. Finally, Section IV-C presents the experimentation of VC_UY on the matrix-matrix multiplication.

A. Matrix Multiplication

Matrix multiplication [46], [47], [48], [49] is an operation that produces a matrix product from two input matrices. Each matrix product entry is the dot product of a row in the first matrix and a column in the second (see equation 2). Matrix multiplication has many applications. It is the basic computational kernel for many algorithms of machine learning systems and recommendation systems. Matrix-vector multiplication is the core kernel of the PageRank algorithm [21], [48]. Matrix multiplication is often a computational bottleneck because generally, matrices are very large with dimensions which can easily reach hundreds, thousands and even millions [46], [47], [48], [49].

Three popular algorithms for matrix multiplication have been proposed in the literature: the iterative algorithm, the recursive algorithm and the Strassen algorithm. Let us consider two square matrices A and B given below.

$$\text{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix},$$

$$\text{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} & \dots & b_{1n} \\ b_{21} & b_{22} & b_{23} & \dots & b_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & b_{n3} & \dots & b_{nn} \end{bmatrix}$$

The matrix multiplication of A and B gives the matrix C obtained by using the equation 2.

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \tag{2}$$

From the previous equation, a simple iterative algorithm can be constructed using loops on the indices i, j from 1 to n. This algorithm takes time on the order of n^3. Using the Divide and Conquer approach, the matrices A, B are partitioned into blocks. Then, the multiplication gives:

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \times \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

The Divide and Conquer approach works for all square matrices whose dimensions are powers of two. For matrices which do not respect this condition (e.g., matrix-vector multiplication), fill the missing rows and columns with zeros. The complexity of the Divide and Conquer approach is the same as the iterative algorithm, i.e. n^3. In fact, this approach requires 8 blocks multiplications to calculate the product matrix which still requires n^3 running time [46], [47], [49].

Another matrix multiplication approach is called Strassen algorithm. The Strassen algorithm [46], [50] is a recursive Divide and Conquer approach. For each recursive call, the input matrices are divided into 4 blocks but only 7 blocks multiplications are needed. Compared to the iterative and the Divide and Conquer approach, the Strassen approach needs only 7 block matrix multiplications, which involves a time complexity of $n^{2.807}$ [46], [47], [49]. When the matrices are large, the execution time of the matrix multiplication can be very long. Since matrix multiplication can be divided into sub-matrix multiplication, this task can be parallelized. The next sections will present how our Volunteer Computing System was built and used for matrix multiplication. All experiments will be performed using the Strassen algorithm.

B. VC_UY: The Volunteer Computing System of the University of Yaounde I

In order to overcome the HPC problems faced by researchers and students from the university of Yaounde I in Cameroon, it was decided to build a VC system named VC_UY. This section presents the summary of the pilot phase in two main points: the recruitment of volunteers and the designing of the VC system.

1) Recruitment of Volunteers: During the volunteer recruitment phase, the master students at the Department of Computer Science of the University of Yaounde I were met. Then, the HPC concepts and the problems encountered in the building of HPC platforms in resource poor settings were explained. Our approach based on volunteering was presented and their participation as volunteers was asked. Ten students were selected from those who generally connect to the Internet at least twice a day and whose computers can be available for computations for at least one hour per day. Table I presents the profile of each volunteer.

TABLE I. Characteristics of Volunteer Computers

RAM	CPU	SWAP	Operating system	Disk space
4GB	Core i3, 2.4GH	1GB	LINUX	50GB
4GB	Core i3, 2.4GH	2GB	WINDOWS	100GB
4GB	Core i4, 2.4GH	2GB	LINUX	100GB
4GB	Core i5, 2.7GH	1.5GB	LINUX	100GB
4GB	Core i5, 2.4GH	1.5GB	LINUX	50GB
4GB	Core i5, 2.4GH	1GB	LINUX	50GB
4GB	Core i3, 2.4GH	2GB	WINDOWS	50GB
4GB	Core i4, 2.4GH	750MB	LINUX	100GB
4GB	Core i4, 2.6GH	1GB	LINUX	100GB
4GB	Core i4, 2.4GH	2GB	LINUX	100GB

2) VC_UY system design: Once the volunteers were recruited, our system was designed as follows: the machine (CPU core i5 and 4 GB of RAM) was used as a server. The VCSoftware[10] was deployed on the server and on each volunteers.

As presented in Section III, the server and volunteers run different modules to exchange information and perform tasks. For the purpose of experimentation, all our source code was written using the Python programming language[11], IPython Application Programming Interface[12] and the Django framework[13]. IPython is an API for parallel and distributed computing. It enables to develop, execute, debug and monitor interactively all types of parallel applications. The architecture of serverManager and volunteersManager modules is completely based on the architecture of IPython API. The Django framework has enabled the implementation of a user-friendly web interface for volunteers (registration, consultation of information about other volunteers, etc.) and the manager (for monitoring using a dashboard).

The serverManager module consists of the front-end developed using the Django framework and the back-end developed using the python programming language and the IPython API. The main feature of the front-end is to support all managements activities (Section III-A) by presenting relevant information on a dashboard. On the back-end side, the IPython API allows the server to listen to the network and collect information on volunteers. For the purpose of experimentation, a script has been written[14] allowing to collect information on volunteers; and a software developed for the visualization of the contribution of each volunteers and the performance of the whole system.

The volunteersManager module consists of the front-end and the back-end. At the front-end, the Django framework presents a dashboard describing all the other volunteers to the volunteer. At the back-end side, the volunteer machine performs the following operations: sends computation to other volunteers; listens to the requests on the network, executes the code, and returns results back to related volunteers; accepts tasks, performs them; collects the results and sends them back to the volunteers; sends profile information to the server. For the purpose of experimentation, scripts have been written (available on github[15]) and permitting to send matrices to volunteers, perform the Strassen matrix multiplication algorithm and collect results.

C. Experimentation of Matrix-Matrix Multiplication on VC_UY

To test the VC_UY Volunteer Computing system, the Strassen matrix-matrix multiplication algorithm was implemented using Python and IPython API. Matrix multiplication was used with different input sizes on one computer (CPU Core i5 and 4GB of RAM) and the whole VC system. If A and B are block-partitioned matrices, the block dimension of the resulting matrix C is determined by considering the number of volunteers and the input block dimensions. A volunteer responsible for each resulting block retrieves all the necessary blocks from A and B to execute a multiplication operation locally. Fig. 2 shows the performance of the execution on one machine. Fig. 3 presents the matrix multiplication on VC_UY system, and Fig. 4 presents the contribution of each volunteers in the computing. Fig. 4 shows that although the volunteers used have close computing power, their contribution to the calculation varies according to their availability.

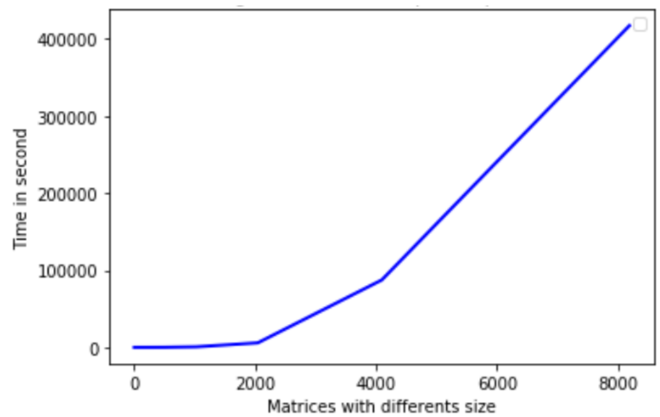


Fig. 2. Experimenting the Strassen algorithm for matrix-matrix multiplication on one machine Core i5 of CPU and 4GB of RAM

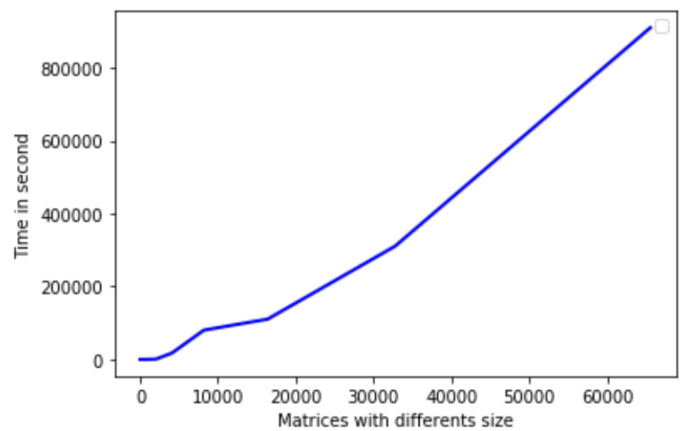


Fig. 3. Experimenting the Strassen algorithm for matrix-matrix multiplication on our Volunteer Computing system

As demonstrated by [25], [26], [51], the experiments conducted in this section shows that Volunteer Computing systems can provide a significant amount of computing power. Then, it is an important option for HPC problems in resource poor settings.

V. Conclusion

This article presented an approach allowing to build a Volunteer Computing system for HPC in resource poor settings. The experiments were made on the mathematical problem of solving matrix multiplication. Volunteers were recruited amongst students at the University of Yaounde I in Cameroon.

[10]https://github.com/admhamza/VC_UY
[11]https://www.python.org/
[12]https://ipyparallel.readthedocs.io/
[13]https://www.djangoproject.com/
[14]https://github.com/admhamza/VC_UY/blob/master/automatisation_collecte_informations.py
[15]https://github.com/admhamza/VC_UY/blob/master/calcul_distribue_dans_un_reseaux.py

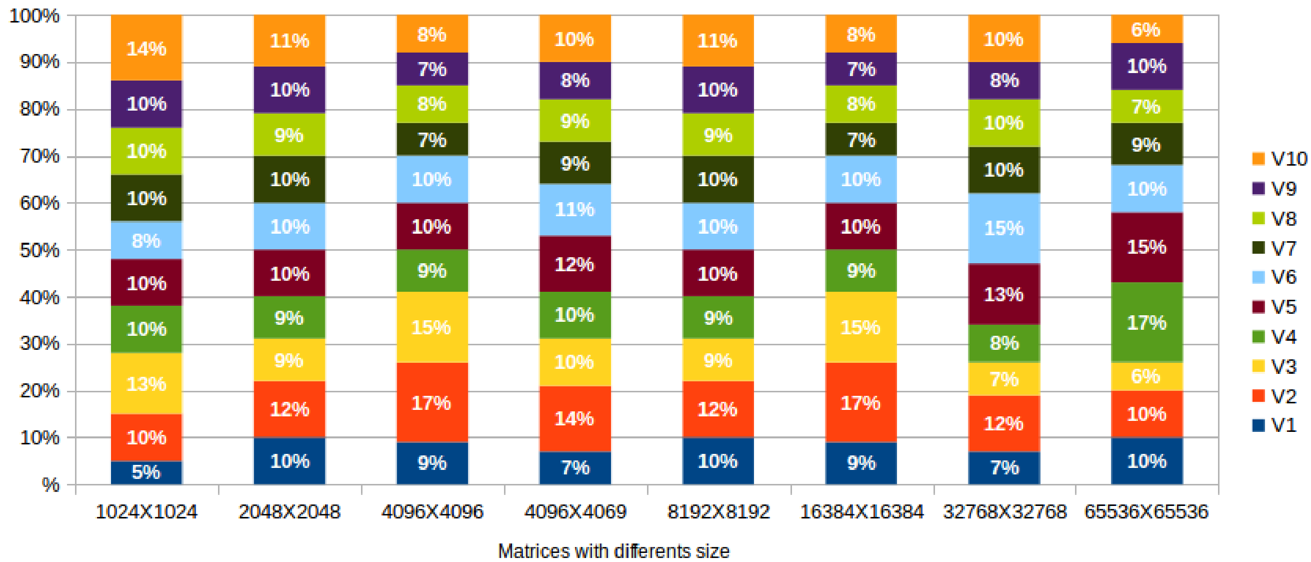


Fig. 4. Contribution of each machine. V_i represents the i^{th} volunteer

Experiments showed that a VC system can be used to enhance HPC in resource poor settings.

This work opens doors for significant possibilities in developing countries where the computing resources are generally limited. Given the success of the experiments, the recruitment of other volunteers from the University of Yaounde I is in progress. The goal being to create a powerful Volunteer Computing system with a maximum number of computers. During the experiment, potential volunteer machines were selected manually according to the data collected. This can be a difficult task if there are hundreds of volunteers registered in the system. Thus, also planned is the exploration and implementation of automatic methods for predicting volunteer machines for a given task.

Acknowledgments

Our gratitude goes to all students who accepted to participate in this project, in particular Mr. Romeo Koati who recruited volunteers and participated to the development of the software.

References

[1] G. S. Almasi and A. Gottlieb. *Highly Parallel Computing*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA, 1989.

[2] Zahid Ansari, Asif Afzal, Moomin Muhiuddeen, and Sudarshan Nayak. Literature survey for the comparative study of various high performance computing techniques. *Int J Comput Trends Technol (IJCTT)*, 27(2):80–86, 2015.

[3] Fatima El Jamiy, Abderrahmane Daif, Mohamed Azouazi, and Abdelaziz Marzak. An effective storage mechanism for high performance computing (hpc). *International Journal of Advanced Computer Science and Applications*, 6(10), 2015.

[4] Zhiwei Xu, Xuebin Chi, and Nong Xiao. High-performance computing environment: a review of twenty years of experiments in China. *National Science Review*, 3(1):36–48, 01 2016.

[5] Ranjit Rajak. A comparative study: Taxonomy of high performance computing (hpc). *International Journal of Electrical & Computer Engineering (2088-8708)*, 8, 2018.

[6] Olaf Schenk, Helmar Burkhart, and Hema Reddy. Towards personalized medicine: High-performance computing in the life sciences. *ERCIM News*, 2008(74), 2008.

[7] Shaoliang Peng. High performance computational biology and drug design on tianhe supercomputers. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016, Shenzhen, China, December 15-18, 2016*, page 7, 2016.

[8] Bertil Schmidt and Andreas Hildebrandt. Next-generation sequencing: big data meets high performance computing. *Drug discovery today*, 22(4):712–717, 2017.

[9] Georg Hager and Gerhard Wellein. *Introduction to High Performance Computing for Scientists and Engineers*. Chapman and Hall / CRC computational science series. CRC Press, 2011.

[10] Daniel A Reed and Jack Dongarra. Exascale computing and big data. *Communications of the ACM*, 58(7):56–68, 2015.

[11] Bo Tang, Zhen Chen, Gerald Hefferman, Shuyi Pei, Tao Wei, Haibo He, and Qing Yang. Incorporating intelligence in fog computing for big data analysis in smart cities. *IEEE Transactions on Industrial informatics*, 13(5):2140–2150, 2017.

[12] Anwar Osseyran and Merle Giles. *Industrial applications of high-performance computing: best global practices*, volume 25. CRC Press, 2015.

[13] Ezell Stephen and Atkinson Robert. *The Vital Importance of High-Performance Computing to U.S. Competitiveness*. INFORMATION TECHNOLOGY INNOVATION FOUNDATION, 2016.

[14] N. R. Adiga, G. Almasi, G. S. Almasi, Y. Aridor, R. Barik, D. Beece, R. Bellofatto, G. Bhanot, R. Bickford, and M. Blumrich et al. An overview of the bluegene/l supercomputer. In *SC '02: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing*, pages 60–60, Nov 2002.

[15] Erich Strohmaier, Hans Werner Meuer, Jack J. Dongarra, and Horst D. Simon. The TOP500 list and progress in high-performance computing. *IEEE Computer*, 48(11):42–49, 2015.

[16] Douglas Eadline. *High Performance Computing For Dummies*. Wiley Publishing, Inc., USA, 5th edition, 2009.

[17] Chee Shin Yeo, Rajkumar Buyya, Hossein Pourreza, Rasit Eskicioglu, Peter Graham, and Frank Sommers. *Cluster Computing: High-Performance, High-Availability, and High-Throughput Processing on a Network of Computers*, pages 521–551. Springer US, Boston, MA, 2006.

[18] Zhe Fan, Feng Qiu, Arie Kaufman, and Suzanne Yoakum-Stover. Gpu cluster for high performance computing. In *Proceedings of the 2004 ACM/IEEE Conference on Supercomputing*, SC '04, pages 47–, Washington, DC, USA, 2004. IEEE Computer Society.

[19] Violeta Holmes and Ibad Kureshi. Developing high performance computing resources for teaching cluster and grid computing courses. *Procedia Computer Science*, 51:1714 – 1723, 2015. International Conference On Computational Science, ICCS 2015.

[20] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. A view of cloud computing. *Commun. ACM*, 53(4):50–58, 2010.

[21] Abhishek Gupta, Laxmikant V Kale, Filippo Gioachin, Verdi March, Chun Hui Suen, Bu Sung Lee, Paolo Faraboschi, Richard Kaufmann, and Dejan Milojicic. The who, what, why and how of high performance computing applications in the cloud. *HP Laboratories Technical Report*, (49), 8 2013.

[22] Ian Foster and Carl Kesselman, editors. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[23] Peter M. Kasson, Daniel L. Ensign, and Vijay S. Pande. Combining molecular dynamics with bayesian analysis to predict and evaluate ligand-binding mutations in influenza hemagglutinin. *Journal of the American Chemical Society*, 131(32):11338–11340, 2009. PMID: 19637916.

[24] Zied TRIFA, Mohamed LABIDI, and Maher KHEMAKHEM. Arabic cursive characters distributed recognition using the dtw algorithm on boinc: Performance analysis. *International Journal of Advanced Computer Science and Applications*, 2(3), 2011.

[25] Erick Lavoie and Laurie Hendren. Personal volunteer computing. In *Proceedings of the 16th ACM International Conference on Computing Frontiers*, CF '19, pages 240–246, New York, NY, USA, 2019. ACM.

[26] Erick Lavoie, Laurie Hendren, Frederic Desprez, and Miguel Correia. Pando: Personal volunteer computing in browsers, 2018.

[27] Oded Nov, David Anderson, and Ofer Arazy. Volunteer computing: A model of the factors determining contribution to community-based scientific research. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 741–750, New York, NY, USA, 2010. ACM.

[28] Haohuan Fu, Junfeng Liao, Jinzhe Yang, Lanning Wang, Zhenya Song, Xiaomeng Huang, Chao Yang, Wei Xue, Fangfang Liu, Fangli Qiao, et al. The sunway taihulight supercomputer: system and applications. *Science China Information Sciences*, 59(7):072001, 2016.

[29] Shurong Tian, Todd Takken, Vic Mahaney, Christopher Marroquin, Mark Schultz, Mark Hoffmeyer, Yuan Yao, Kevin O'Connell, Anil Yuksel, and Paul Coteus. Summit and sierra supercomputer cooling solutions. *IBM Journal of Research and Development*, 2019.

[30] James R Kozloski, Timothy M Lynar, Mark D Nelson, and John M Wagner. Energy efficient supercomputer job allocation, 2018. US Patent 10,025,639.

[31] Rajkumar Buyya. *High Performance Cluster Computing: Architectures and Systems*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1999.

[32] Zvi Tannenbaum and Dean E Dauger. Cluster computing, 2019. US Patent 10,333,768.

[33] Minakshi Tripathy and C.R. Tripathy. On a virtual shared memory cluster system with virtual machines. *International Journal of Computer and Electrical Engineering*, 3:754–761, 01 2011.

[34] Mohammad Moghadasi, Seyed Majid Mousavi, and Gábor Fazekas. Cloud computing auditing. *International Journal of Advanced Computer Science and Applications*, 9(12), 2018.

[35] Aferdita Ibrahimi. Cloud computing: Pricing model. *International Journal of Advanced Computer Science and Applications*, 8(6), 2017.

[36] Babur Hayat Malik, Jazba Asad, Sabila Kousar, Faiza Nawaz, Zainab, Farania Hayder, Sehresh Bibi, Amina Yousaf, and Ali Raza. Cloud computing adoption in small and medium- sized enterprises (smes) of asia and africa. *International Journal of Advanced Computer Science and Applications*, 10(5), 2019.

[37] Hans-Joachim Bungartz. Sparse grids and their impact on hpc and big data. In *Kolloquiumsvortrag*, 2019.

[38] Tessema M Mengistu and Dunren Che. Survey and taxonomy of volunteer computing. *ACM Computing Surveys (CSUR)*, 52(3):1–35, 2019.

[39] Bruno Donassolo, Henri Casanova, Arnaud Legrand, and Pedro Velho. Fast and scalable simulation of volunteer computing systems using simgrid. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, HPDC '10, pages 605–612, New York, NY, USA, 2010. ACM.

[40] David P. Anderson, Jeff Cobb, Eric Korpela, Matt Lebofsky, and Dan Werthimer. Seti@home: An experiment in public-resource computing. *Commun. ACM*, 45(11):56–61, 2002.

[41] A. L. Beberg, D. L. Ensign, G. Jayachandran, S. Khaliq, and V. S. Pande. Folding@home: Lessons from eight years of volunteer distributed computing. In *2009 IEEE International Symposium on Parallel Distributed Processing*, pages 1–8, May 2009.

[42] Wei Li and William W Guo. The optimization potential of volunteer computing for compute or data intensive applications. *Journal of Communications*, 14(10), 2019.

[43] Harold E Castro. Capacity of desktop clouds for running hpc applications: A revisited analysis. In *Applied Informatics: Second International Conference, ICAI 2019, Madrid, Spain, November 7–9, 2019, Proceedings*, volume 1051, page 257. Springer Nature, 2019.

[44] George Coulouris, Jean Dollimore, Tim Kindberg, and Gordon Blair. *Distributed Systems: Concepts and Design*. Addison-Wesley Publishing Company, USA, 5th edition, 2011.

[45] Maarten Van Steen and Andrew S Tanenbaum. *Distributed systems*. Maarten van Steen Leiden, The Netherlands, 2017.

[46] Qingshan Luo and John B. Drake. A scalable parallel strassen's matrix multiplication algorithm for distributed-memory computers. In *Proceedings of the 1995 ACM Symposium on Applied Computing*, SAC '95, pages 221–226, New York, NY, USA, 1995. ACM.

[47] Chandan Misra, Sourangshu Bhattacharya, and Soumya K. Ghosh. Stark: Fast and scalable strassen's matrix multiplication using apache spark, 2018.

[48] M. Son and K. Lee. Distributed matrix multiplication performance estimator for machine learning jobs in cloud computing. In *2018 IEEE 11th International Conference on Cloud Computing (CLOUD)*, pages 638–645, July 2018.

[49] Zhen Xie, Guangming Tan, Weifeng Liu, and Ninghui Sun. Ia-spgemm: An input-aware auto-tuning framework for parallel sparse matrix-matrix multiplication. In *Proceedings of the ACM International Conference on Supercomputing*, ICS '19, pages 94–105, New York, NY, USA, 2019. ACM.

[50] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009.

[51] Travis Desell. Developing a volunteer computing project to evolve convolutional neural networks and their hyperparameters. pages 19–28, 10 2017.

Search Space of Adversarial Perturbations against Image Filters

Dang Duy Thang[1], Toshihiro Matsui[2]
Institute of Information Security,
Yokohama, Japan

***Abstract*—The superiority of deep learning performance is threatened by safety issues for itself. Recent findings have shown that deep learning systems are very weak to adversarial examples, an attack form that was altered by the attacker's intent to deceive the deep learning system. There are many proposed defensive methods to protect deep learning systems against adversarial examples. However, there is still lack of principal strategies to deceive those defensive methods. Any time a particular countermeasure is proposed, a new powerful adversarial attack will be invented to deceive that countermeasure. In this study, we focus on investigating the ability to create adversarial patterns in search space against defensive methods that use image filters. Experimental results conducted on the ImageNet dataset with image classification tasks showed the correlation between the search space of adversarial perturbation and filters. These findings open a new direction for building stronger offensive methods towards deep learning systems.**

Keywords*—Deep neural networks; image filters; adversarial examples; image classification

I. Introduction

Over the past decade, there has been the rise of deep learning in many tasks such as computer vision [1], automatic driving [2], natural language processing [3], and so on. Deep learning models are designed based on an assumption of inputs and outputs distribution being benign. This leads to when training deep learning models, we only focus to fine-tune the weights, parameters or the number of nodes and hidden layers while setting aside the validity of data. This has created a security issue against deep learning systems. Szegedy et al. [4] explored that deep neural networks are at risk of attacks from adversarial example attacks. Afterward, many research work on technologies that delude AI models has gradually become a hot spot, and researchers have continued to propose new methods of attack and defense. Adversarial attacks have been regularly adapted in both research and commerce. In the computer vision area, many adversarial attacks are proposed in image classification [5], [6], [7], [8], and object detection [9]. There are also many researches work on the adversarial example in text [10], [11], [12], [13]. In the physical world attack, Kurakin et al. [14] first exposed that hazards of adversarial examples. They use an application of Tensor-Flow Camera Demo to capture original images. After that they use Google Inception V3 [15] for classifying those images. The implementation has been shown that a large portion of the image has been misclassified even when observed via the camera lens. Eykholt Kevin et al. [16] invented a new method based on [7] and [17] to create robust adversarial perturbation in the real world. They indicated variation in view angles, distance, and resolution are almost defeated by the robust adversarial examples in physical settings. The proposed algorithm used a term as RP_2 for Robust Physical Perturbations, which was used to craft adversarial examples for road sign recognition systems that perform a high deceiving rate in an efficient setting. And many physical adversarial attacks are proposed in face recognition [18], machine vision [19], and road sign recognition [20]. In the cyberspace security field, there are adversarial attacks in cloud service [21], malware detection [22], [23], and network intrusion detection [24]. Besides the attack methods, many defensive approaches have been proposed and they can be branched into four main categories include adversarial training, denoising, transformation and compression. Szegedy et al. [4] used adversarial examples to train an AI model with the ground truth labels, and it made that model more robust. Goodfellow et al. [5] also used the adversarial training strategy to improve the classification rate on adversarial examples with the MNIST dataset. Tramèr et al. [25] combined the adversarial examples created from many different AI models to increase the robustness of those models. [26], [27] proposed new methods based on the image transformation to reduce the misclassification rate of an AI model. [28], [29] assumed almost adversarial examples are created in the high-frequency domain and they proposed the new method based on image filters to remove the adversarial perturbations. Das et al. [30] introduced a defensive method based on JPEG compression to deceive FGSM [5] and DeepFool [31] attacks. However, newer adversarial attacks such as Carlini&Wagner attacks [7] overcame these compression defensive strategies.

Our Contributions. In this work, we investigate the search space of adversarial perturbation. A challenge in the process of understanding the effects of adversarial noises is very limited so far. How to determine the available space of adversarial noises is very important. Understanding and identifying this space will help us develop better protection systems for deep learning against adversarial examples.

We describe our main contributions of this research as below:

- We have recapped the numerous adversarial defensive and attack methods. Moreover, we have provided a perceptive review of these current methods.
- We discovered the close relationship between search space of adversarial perturbation and image filters.
- Our research opens up a new perspective on creating stronger and more effective attacks on deep learning systems.

Paper outlines. The remainder of our paper is described as follows. Section II introduces the literature review and the background of adversarial examples. Section III describes our approach on search space of adversarial examples, and Section IV demonstrate our implementation and evaluation results. Section V summaries our work.

II. Literature Review and Background

A. Literature review

In this work, we focus on the relation between feasible space of adversarial perturbation and defensive methods based on frequency domain. So we make a literature review on these defensive methods in this section. Eliminating the adversarial features and retaking the classification rate have been considered in many works. Xu et al. [32] proposed a new defensive approach by using the feature squeezing strategies to remove the adversarial features. There are two key ideas in [32]. The first one considered the bit depth in an input image. By increasing or reducing the bit depth of image, the method removed some adversarial features. The second one used the median filter to defeat the adversarial features. However, [32] required a range of thresholds to separate between adversarial and legitimate features. So the selection of a relevant threshold for a specific dataset or setting is a nontrivial task and it is heuristic. Dang et al. [28] proposed a detection system for automatically identifying adversarial examples with the image filters (Gaussian, Median filter). The system doesn't require to setup any threshold for distinguishing adversarial and benign images. However, there is unclear how the system is able to suffer the stronger and new adversarial attacks. Our paper shows that Gaussian blurring only works well on the small adversarial perturbation, and it is futile to larger and stronger adversarial perturbation.

B. Background

1) Convolutional Neural Networks: Convolutional Neural Networks (CNNs) are designed to learn the important features from the training dataset to match them with the given labels. CNNs are used in many areas [1], [3] and provided open-source [15]. CNNs include multilayers with many operations to process signals from a lower layer to a higher layer in hierarchy architecture. In this research, we emphasize in image classification task so we only cover the brief fundamentals in this area. In an image classification task, CNNs process an input data x and try to figure out the best matching output label y from a set of labels Y. The structure of a CNN can be described as shown in Table I. The layers are described in a top-down order from input to output. We can see for this CNN network, the input is a color image of size 299×299. The first layer is a convolutional layer whose kernel size is 3×3 with a stride of 2. The next convolutional layers also use the same kernel size with a difference with the number of kernels as well as stride. In an inception network, it appears layers called inception layers. These inception layers are different from convolutional layers in that they combine several different kernel sizes at once to extract more important features. The inception layer can also be called inception filters. The last adjacent layer is the logits layer before the softmax function is implemented to calculate the probability for each output label corresponding to the input.

TABLE I. Google Inception Architecture [15]

layer	patch/stride or note	input
conv	3 × 3/2	299 × 299 × 3
conv	3 × 3/1	149 × 149 × 32
conv padded	3 × 3/1	147 × 147 × 32
pool	3 × 3/2	147 × 147 × 64
conv	3 × 3/1	73 × 73 × 64
conv	3 × 3/2	71 × 71 × 80
conv	3 × 3/1	35 × 35 × 192
3 × Inception	Inception filters	35 × 35 × 288
5 × Inception	Inception filters	17 × 17 × 768
2 × Inception	Inception filters	8 × 8 × 1280
pool	8 × 8	8 × 8 × 2048
linear	logits	1 × 1 × 2048
softmax	classifier	1 × 1 × 1000

2) Adversarial Attacks: Adversarial examples are defined as malicious patterns created by the slightly modified aim to fool an AI model but indistinguishable from humans.

FGSM (Fast Gradient Sign Method) was proposed by Ian Goodfellow et al. [5]. In a normal training process, the input and output data distributions are assumed as fixed and unchangeable, so there are only trainable parameters and weights that are fine-tuned respect to a loss objective function between input x and label y. [5] used a very simple idea to reverse that normal process when they fine-tuned input data distribution respect to a new loss objective function between new sample x^{adv} with new specific label y^{adv}:

$$x^{adv} = x - \beta \cdot sign(\nabla_x Loss(x^{adv}, y^{adv})) \quad (1)$$

where β denotes the perturbation size to create an adversarial example x^{adv} from a legitimate input x. From a legitimate input x, FGSM looks for the best adversarial perturbation β to add into x to create a new image x^{adv}. The value of β has to satisfy two requirements include the magnitude of β is as small as possible and respect to the loss objective function between (x,y). For the first requirement, the magnitude of β is smaller, x^{adv} is more similar as x but the convergence rate of the algorithm 1 is slower, while the bigger β makes x^{adv} is more different from the x but the FGSM algorithm converges faster. For the second requirement, the loss objective function between (x,y) is maximized and $Loss(x^{adv}, y^{adv})$ is minimized. Because the total of probabilities of output is equal to one, so the algorithm 1 only needs to consider to minimize $Loss(x^{adv}, y^{adv})$.

In this paper, we use the FGSM [5] method with l-norm optimization as a baseline to conduct assessments of the possible value areas of β during the creation of adversarial examples. Our attack method is based on a white-box attack where victim AI model information is known in advance and can be accessed.

3) Defensive approaches: There are many methods of protection that have been proposed. The typical strategy is adversarial training [4], [33], [25]. The idea of this strategy of protection is that the AI models will be trained with adversarial examples and ground truth labels. With the assumption that the more AI models are learned, the more accurate they will regain and the more likely it will be to misidentify adversarial examples. However, the major drawback of the adversarial training method is that it takes a lot of time to create adversarial examples and training time for AI models. In addition, this method does not guarantee resistance to new adversarial examples created by other methods than those created by the previous method. Other defensive methods that are often investigated to be pre-processing data. These defensive methods include preprocessing methods based on image transformation [26], [27], filter [28], [29] or compression [30]. Those methods of defense have very impressive results in helping AI systems identify which input is adversarial or legitimate. One of the defenses which also attracts high attention is gradient masking. The adversarial attack methods are largely based on gradient calculations to optimize the loss objective function when creating adversarial examples. For that reason, the idea of hiding the gradient value was proposed. [25] proposed a gradient masking method based on smoothing the gradient gradients that made the global optimal calculation based on gradient slope is more difficult. Author in [34] uses another strategy that is distillation synthesized from different models to create a stronger model against adversarial examples.

III. Search Space of Adversarial Perturbation

A. Search Space on Attacking Phase

One of the important factors in the process of creating adversarial examples is the adversarial perturbation coefficient β. However, how to find out the optimal value of β and its relationship to the currently most powerful defense methods in relation to image filter [28] is unclear. That is the purpose of this study. In this research, we investigate on a white-box attack in creating adversarial examples. This is the setting defined as the attacker can access and use the AI model parameters for conducting an attack pattern. This is possible because currently, the most powerful AI models in image classification tasks are open-source. Many attack methods have been proposed, but most of them rely on FGSM for development, generally, we also use FGSM for creating adversarial examples. One thing to note, it is possible to classify adversarial attacks into two different types based on the purpose of the attacker include non-targeted and targeted attacks. The non-targeted attack is defined as the attacker only focuses on maximize the loss function of (x, y) in order to deceive the AI system. Meanwhile, a targeted attack is defined as the attacker wants to trick the AI system into a misclassifying new pattern in an intentional label rather than merely misidentifying it. Because of this, targeted attacks are more commonly used than non-targeted attacks and we also use it in attacking phases. Our main purpose to decide the size of adversarial perturbation, it means the search space of adversarial perturbation. We consider the norm operation to determine the size of the adversarial noises. Mathematically, the norm operation is used to calculate the distance, or the length of the vectors or the matrixes according to element-wise. The bigger the norm value, the bigger the difference between vectors or matrices and vice versa. Formally, the l_p-norm of vector x is defined as: $\|x\|_p = \sqrt[p]{\sum_i |x_i|^p}$, where $p \in \mathbb{R}$. This is a p^{th}-root of a summation of all elements to the p^{th} power is what we call a norm. The important point is even though every l_p-norm is all looked very similar to each other, their mathematical properties are very different and thus their application is completely different when we use to create the adversarial examples. In this work, we consider three common norm methods: l_1-norm, l_2-norm, and l_∞-norm for evaluating the size of the search space of adversarial perturbation.

$\mathbf{l_1}$-norm. We define x_{true} as the original input vector, l_1-norm of x_{true} is defined as:

$$\|x_{true}\|_1 = \sum_i \left| x_{true}^{(i)} \right| \tag{2}$$

This norm is also well-known as the Manhattan norm and it is one of very common norm operations.

$\mathbf{l_2}$-norm. is the most popular norm and also known as the Euclidean norm. The l_2-norm and other norms are equivalent in the sense that all of them are defined in the same topology. The l_2-norm is defined as:

$$\|x_{true}\|_2 = \sqrt{\sum_i (x_{true}^{(i)})^2} \tag{3}$$

We use the l_2-norm to measure the difference between two vectors x_{true} and x_{adv}, the l_2-norm is re-defined:

$$\|x_{true} - x_{adv}\|_2 = \sqrt{\sum_i (x_{true}^{(i)} - x_{adv}^{(i)})^2} \tag{4}$$

where x_{adv} defines the adversarial example.

$\mathbf{l_\infty}$-norm. The l_∞-norm is defined as equation below:

$$\|x_{true}\|_\infty = \sqrt[\infty]{\sum_i (x_{true}^{(i)})^\infty} \tag{5}$$

Let consider the vector x, if $x^{(i)}$ is each element in vector x, from the property of the infinity itself, we have: $x_i^\infty \approx x_k^\infty \forall i \neq k$, then $\sum_i x_i^\infty = x_k^\infty$. And we have $\|x\|_\infty = \sqrt[\infty]{\sum_i x_i^\infty} = \sqrt[\infty]{x_k^\infty} = |x_k|$. Now we have simple definition of l_∞-norm as: $\|x\|_\infty = max(|x_i|)$.

So our attack phase is denoted as Algorithm 1 by using FGSM. Where x_{true} defines the original input, x_{adv} is adversarial example, y_{true} defines the ground-truth label, y_{adv} is an adversarial label, f is the activation function of machine learning model, β is the maximum adversarial value, l_i defines the norm. For crafting adversarial example, we set a learning rate lr is equal to 0.01, the number of iteration is 500 times.

B. Filter Methods

Most adversarial attack methods look for the optimal values of adversarial perturbation respect to loss objective function to modify the original image. Therefore, the pixels that are incidentally edited are located in the high-frequency domain. Therefore current protection methods based on image filters have proved very effective in eliminating these adversarial noises. However, in order to better understand the search space of this adversarial perturbation and the ability to resist image filters, we studied the two most common image filters, the Gaussian and the Median filter. Mathematically, a Gaussian filter modifies the input image by calculating a convolution the area of a specific image area with a Gaussian function; this transformation is also known as the Weierstrass transform. The area of convolution is often called kernel size and is usually 3x3 or 5x5. When using a Gaussian filter, the kernel window will move across the surface of the input image and compute the kernel window that corresponds to the image area being

Algorithm 1: Crafting Adversarial Examples with l-norm optimization

input : x_{true}, y_{true}, y_{adv}, f, β, l_i
output : x_{adv}
parameter: lr = 0.01, iterations = 500

$x_{adv} \leftarrow x_{true}$ `// initial adversarial example`
$\delta \leftarrow \vec{0}$ `// initial adversarial perturbation`
$it \leftarrow 1$ `// initial iteration loop`
while $\delta < \beta$ *and* $f(x_{adv}) \neq y_{adv}$ *and* $it <= iterations$ **do**
  $x_{adv} \leftarrow x_{true} - \delta \cdot sign(\bigtriangledown Loss(y_{adv}|x_{adv}))$
  $\delta \leftarrow$ norm(l_i)
  maximize $Loss(y_{adv}|x_{adv})$ respect to δ
  $\delta \leftarrow clip(x_{adv}, x - \beta, x + \beta)$
  $it \leftarrow it + 1$
end
return x_{adv}

processed. The second image filter to be considered in this research is the median filter. This is a very common filter used to highlight the edges of an image. The Median filter also uses kernel windows that move across the input image surface. However, the median filter processes that area simply by finding the median value of the image area being processed, then replacing that median value in the pixel position in the center of the windows kernel while preserving the pixel values in neighbors. Our filtering system proceeds by Algorithm 2, where x defines the input image, φ denotes the kernel sizes, f is a machine learning function that computes the predicted label with the highest probability, y_{true} defines the ground truth label, y_{adv} defines the adversarial label, and s is the filter function. Output are the probabilities of the ground truth label (p_{true}) and the adversarial label (p_{adv}).

Algorithm 2: Image Filters on input for image classification task

input : $x_{true}, s, f, y_{true}, y_{adv}$
output : p_{true}, p_{adv}
parameter: $\varphi = [(3 \times 3); (5 \times 5)]$

for i *in* φ **do**
  $x_{filtered} \leftarrow s(x, i)$
  $P \leftarrow f(x_{filtered})$
  $p_{true} \leftarrow P(y_{true})$
  $p_{adv} \leftarrow P(y_{adv})$
end

IV. Implementation and Results

A. Datasets and AI model

The target AI model that we use in the implementation is Google Inception V3 [15] that was trained with 1,000 common categories in the ImageNet [35] dataset. Our attacking phase is a white-box attack setting and a targeted label is "street sign" label. We use FGSM with l_1-norm, l_2-norm, and l_∞-norm to craft adversarial images.

B. Results

Intuitively, because of the copyright issue of the ImageNet dataset, we use our own images (include pictures of vending machine, computer mouse and keyboard) for analysis. We randomly selected targeted labels for the creation of adversarial images. By using the FGSM method in combination with l_1-norm, l_2-norm, and l_∞-norm, from each original image we create three different adversarial images.

Fig. 1 shows the probabilities of the original vending machine label when the input is an vending machine image. Fig. 2 shows the probabilities of the adversarial label with

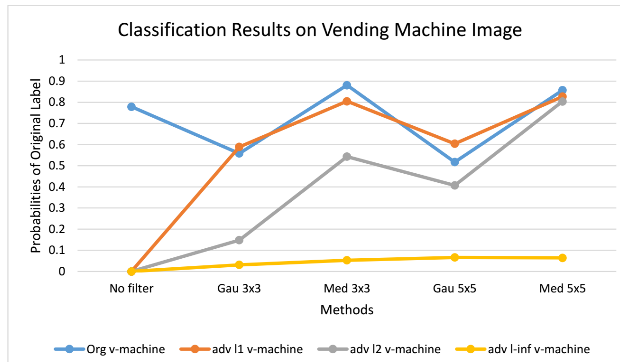


Fig. 1. Classification Results on Vending Machine Image with observation on the probabilities of Original Label

vending machine input. Fig. 3 shows the observations on the

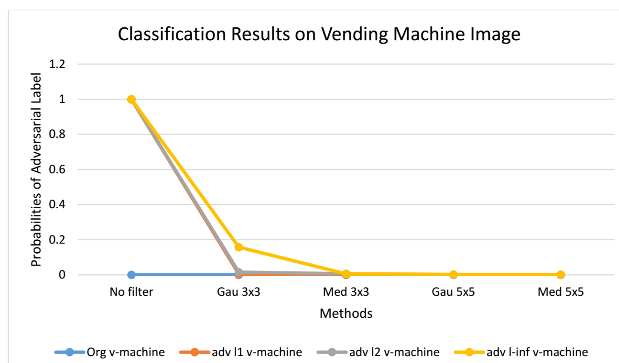
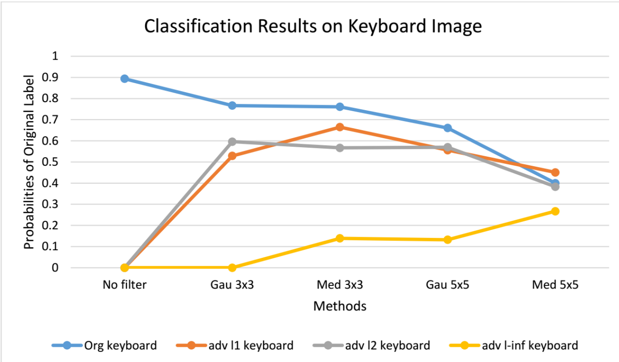


Fig. 2. Classification Results on Vending Machine Image with observation on the probabilities of Adversarial Label

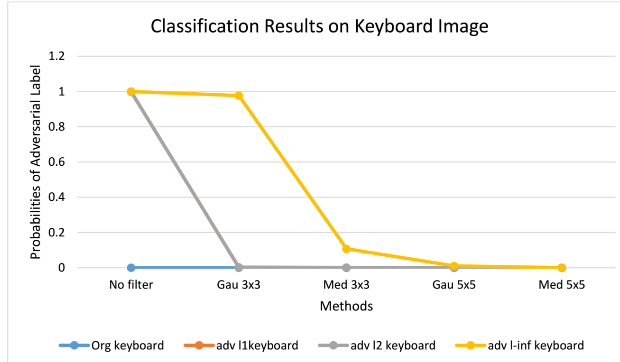
images of computer mouse and keyboard.

Fig. 4 shows the results of creating adversarial images from the original image of the vending machine. We find that the deep learning system is easily fooled with adversarial images. In addition, we intuitively observe that adversarial images created with l_1-norm and l_2-norm are harder to detect than l_∞-norm.

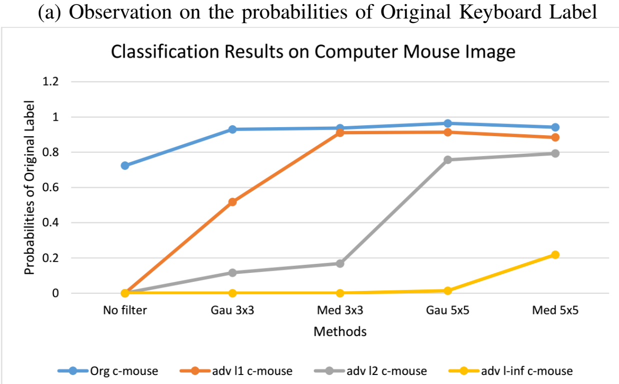
Fig. 5 shows the experimental results when we use the image filters method on the original image of the vending machine. We find that the Gaussian filter reduces classification accuracy more than the median filter. Especially in the case of the median with size filter 3×3 and 5×5, the classification results are better than the original image.



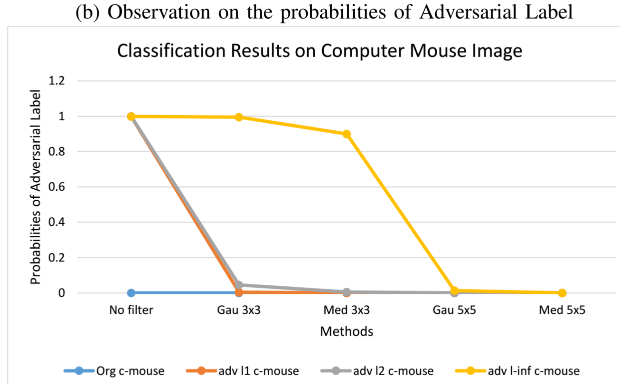
(a) Observation on the probabilities of Original Keyboard Label



(b) Observation on the probabilities of Adversarial Label



(c) Observation on the probabilities of Original Computer Mouse Label



(d) Observation on the probabilities of Adversarial Label

Fig. 3. Classification Results on Keyboard and Computer Mouse Images

Similar to the original image, we also apply image filter methods to adversarial images. Fig. 6 shows classification results on adversarial images created by the FGSM method in combination with l_1-norm. Fig. 7 illustrates classification results on adversarial images created by the FGSM method in combination with l_2-norm. We observed that Gaussian kernel size 3x3 could not restore identity to ground truth label on adversarial image with l_2-norm. The probability for vending machine label is only 14.8%. Meanwhile, the median filter still works effectively in removing adversarial noises. Fig. 8 shows classification results on adversarial images created by the FGSM method in combination with l_∞-norm. We observed that Gaussian kernel size 3×3 could not eliminate the effect of adversarial noise with l_∞-norm on deep learning system classification. Gaussian 5×5 gives better results, but the label with the highest probability of identification is "tabacco shop". The Median filter removes adversarial noises but cannot help the deep learning system correctly identify ground truth labels.

Table II shows experimental results on vending machine (v-machine), computer mouse (c-mouse) and keyboard sets. This result shows us a large correlation between norm operations in search space of adversarial examples. It is clear that for the l_∞-norm, the Gau (3×3, 5×5) and median (3×3) methods are more difficult to completely eliminate adversarial noises based on the l_1 and l_2 norm. Median (5×5) still proved superior in removing adversarial noises in all settings.

V. Conclusion

In this study, we focus on investigating the connection between the search space of adversarial examples and the defense based on the frequency domain. Our empirical results demonstrate that the FGSM method in combination with l_∞-norm produces the strongest adversarial examples. In this case, both the Gaussian and the Median filters are unable to restore identification to the ground truth label. However, when using l_∞-norm to create adversarial examples, we also significantly reduce the quality of the original image compared to using l_1 and l_2 norm. In terms of similarities with the original image, l_1 and l_2 norm produce much better adversarial examples than l_∞ norm.

Acknowledgment

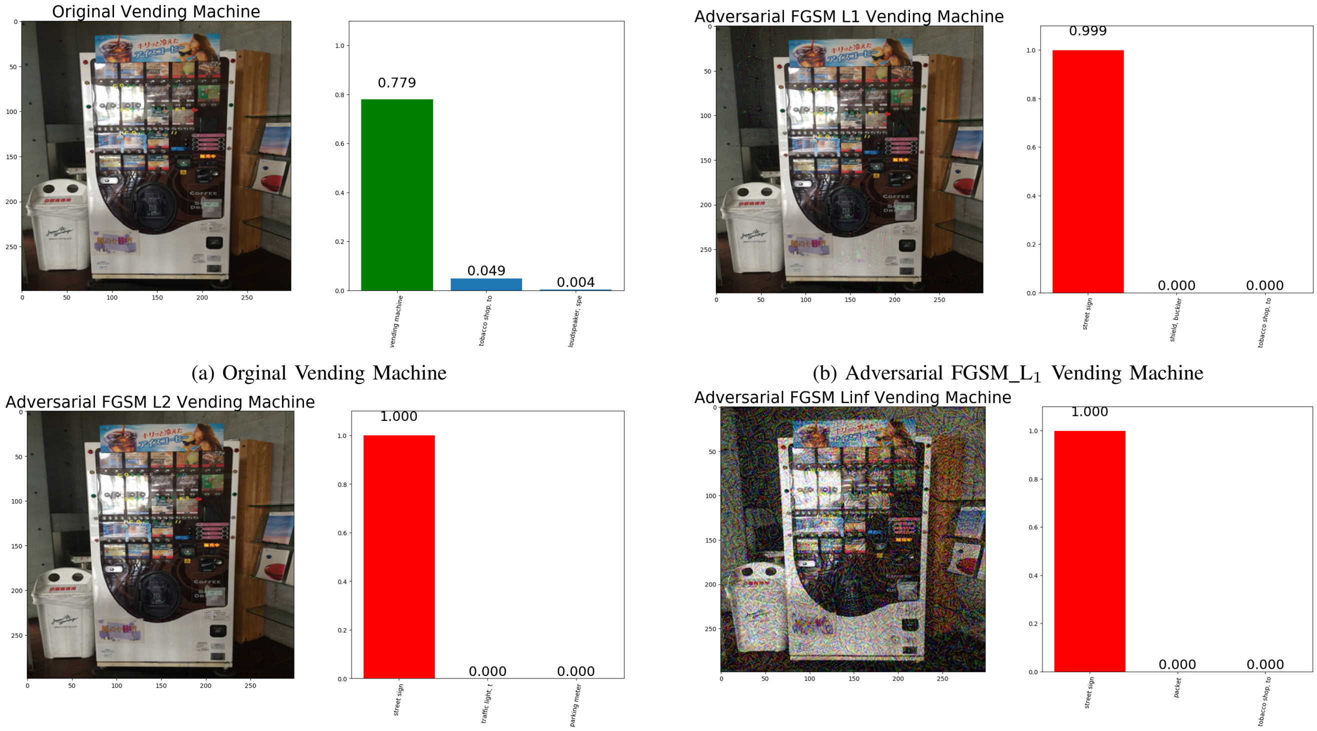
We would like to thank Professor Akira Otsuka for the valuable suggestions on this research. This work was supported by the Iwasaki Tomomi Scholarship.

References

[1] A. Ioannidou, E. Chatzilari, S. Nikolopoulos, and I. Kompatsiaris, "Deep learning advances in computer vision with 3d data: A survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, p. 20, 2017.

[2] D. Watzenig and M. Horn, *Automated driving: safer and more efficient future driving*. Springer, 2016.

[3] J. Ray, O. Johnny, M. Trovati, S. Sotiriadis, and N. Bessis, "The rise of big data science: A survey of techniques, methods and approaches in the field of natural language processing and network theory," *Big Data and Cognitive Computing*, vol. 2, no. 3, p. 22, 2018.



(a) Orginal Vending Machine

(b) Adversarial FGSM_L_1 Vending Machine

(c) Adversarial FGSM_L_2 Vending Machine

(d) Adversarial FGSM_L_∞ Vending Machine

Fig. 4. Adversarial Vending Machine (targeted class: Street Sign)

TABLE II. IMPLEMENTATION RESULTS

Input	No filter		Gau 3x3		Med 3x3		Gau 5x5		Med 5x5	
	OL	AL	OL	AL	OL	AL	OL	AL	OL	AL
Org v-machine	0.779	0	0.558	0	0.881	0	0.517	0	0.857	0
adv l_1 v-machine	0	0.999	0.589	0.004	0.805	0.001	0.604	0.001	0.827	0
adv l_2 v-machine	0	1	0.148	0.015	0.543	**0.006**	0.407	0.001	0.804	0
adv l_∞ v-machine	0	1	0.031	**0.157**	0.053	0.005	0.066	**0.002**	0.064	**0.001**
Org keyboard	0.894	0	0.767	0	0.761	0	0.661	0	0.4	0
adv l_1 keyboard	0	0.999	0.529	0.002	0.665	0	0.556	0	0.451	0
adv l_2 keyboard	0	0.999	0.596	0.002	0.567	0.001	0.57	0	0.383	0
adv l_∞ keyboard	0	1	0	**0.977**	0.139	**0.107**	0.132	**0.01**	0.267	0
Org c-mouse	0.724	0	0.93	0	0.937	0	0.964	0	0.924	0
adv l_1 c-mouse	0	0.999	0.518	0.004	0.911	0	0.914	0	0.884	0
adv l_2 c-mouse	0	1	0.116	0.045	0.168	0.005	0.757	0	0.793	0
adv l_∞ c-mouse	0	0.999	0	**0.995**	0	**0.9**	0.014	**0.013**	0.218	0

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations ICLR*, 2014.

[5] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations ICLR*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[6] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[7] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (S&P 2017)*. IEEE, 2017, pp. 39–57.

[8] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM, 2017, pp. 15–26.

[9] C. Xiao, R. Deng, B. Li, F. Yu, M. Liu, and D. Song, "Characterizing adversarial examples based on spatial consistency information for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 217–234.

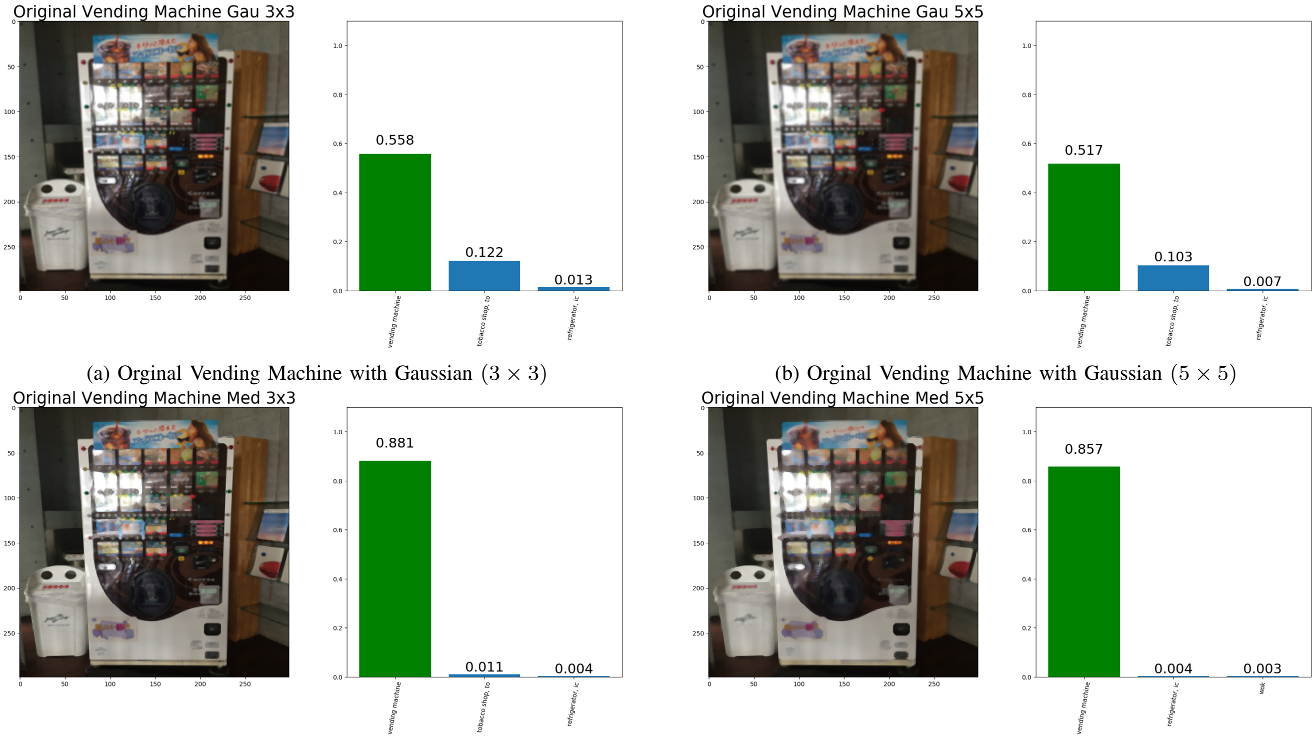
[10] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," in *International Conference on Learning Representations ICLR*, 2017.

[11] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "Hotflip: White-box adversarial examples for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 31–36.

[12] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 50–56.

[13] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "Textbugger: Generating adversarial text against real-world applications," in *26th Annual Network and Distributed System Security Symposium*, 2019.

[14] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the



(a) Orginal Vending Machine with Gaussian (3×3)

(b) Orginal Vending Machine with Gaussian (5×5)

(c) Orginal Vending Machine with Median (3×3)

(d) Orginal Vending Machine with Median (5×5)

Fig. 5. Original Vending Machine with Image Filters

physical world," in *International Conference on Learning Representations ICLR*, 2017.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[16] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2018.

[17] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *International Conference on Learning Representations ICLR*, 2017.

[18] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 1528–1540.

[19] M. Melis, A. Demontis, B. Biggio, G. Brown, G. Fumera, and F. Roli, "Is deep learning safe for robot vision? adversarial examples against the icub humanoid," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 751–759.

[20] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[21] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, 2017, pp. 506–519.

[22] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial perturbations against deep neural networks for malware classification," *arXiv preprint arXiv:1606.04435*, 2016.

[23] H. S. Anderson, J. Woodbridge, and B. Filar, "Deepdga: Adversarially-tuned domain generation and detection," in *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security*. ACM, 2016, pp. 13–21.

[24] C.-H. Huang, T.-H. Lee, L.-h. Chang, J.-R. Lin, and G. Horng, "Adversarial attacks on sdn-based deep learning ids system," in *International Conference on Mobile and Wireless Technology*. Springer, 2018, pp. 181–191.

[25] A. Kurakin, D. Boneh, F. Tramèr, I. Goodfellow, N. Papernot, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *International Conference on Learning Representations ICLR*, 2018.

[26] T. Dang and T. Matsui, "Image transformation can make neural networks more robust against adversarial examples," *CoRR*, vol. abs/1901.03037, 2019. [Online]. Available: http://arxiv.org/abs/1901.03037

[27] D. D. Thang and T. Matsui, "A label-based approach for automatic identifying adversarial examples with image transformation," in *2019 Seventh International Symposium on Computing and Networking (CANDAR)*. IEEE, Nov 2019, pp. 112–120.

[28] D. D. Thang and T. Matsui, "Automated detection system for adversarial examples with high-frequency noises sieve," in *International Symposium on Cyberspace Safety and Security*. Springer, 2019, pp. 348–362.

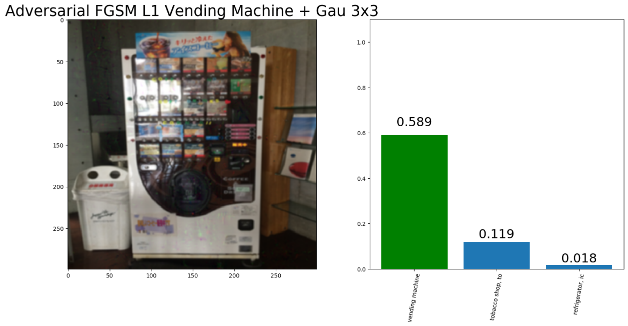
[29] D. D. Thang, T. Kondo, and T. Matsui, "A label-based system for detecting adversarial examples by using low pass filters," *Computer Security Symposium 2019 (CSS2019)*, pp. 1356–1363, 2019.

[30] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau, "Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression," *arXiv preprint arXiv:1705.02900*, 2017.

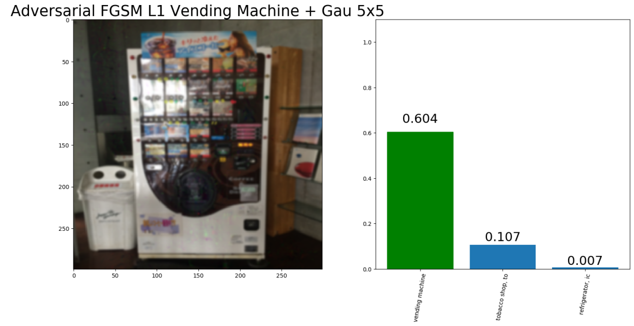
[31] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition CVPR*, 2016, pp. 2574–2582.

[32] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21*, 2018.

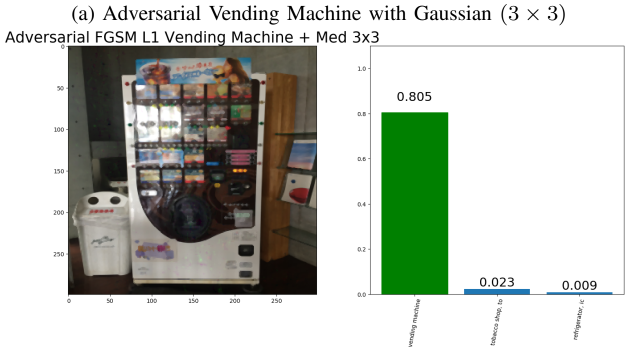
[33] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learn-



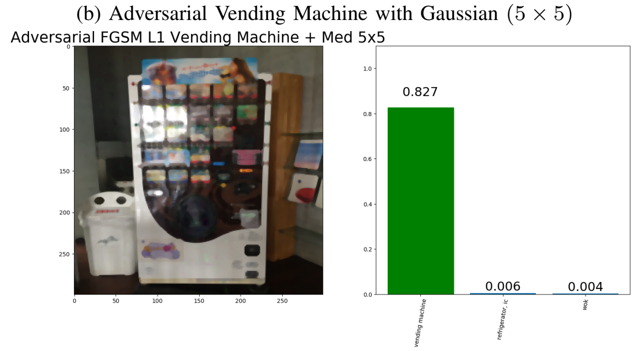
(a) Adversarial Vending Machine with Gaussian (3×3)



(b) Adversarial Vending Machine with Gaussian (5×5)



(c) Adversarial Vending Machine with Median (3×3)



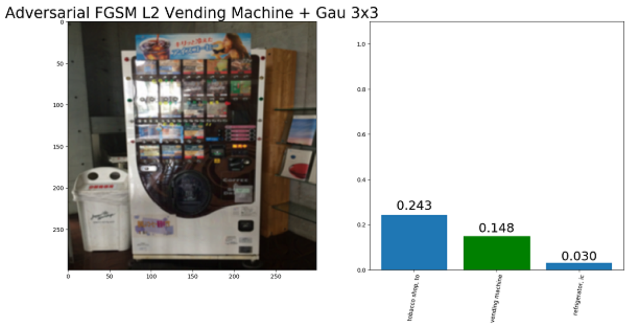
(d) Adversarial Vending Machine with Median (5×5)

Fig. 6. Adversarial FGSM L_1 Vending Machine (targeted class: Street Sign) with Image Filters

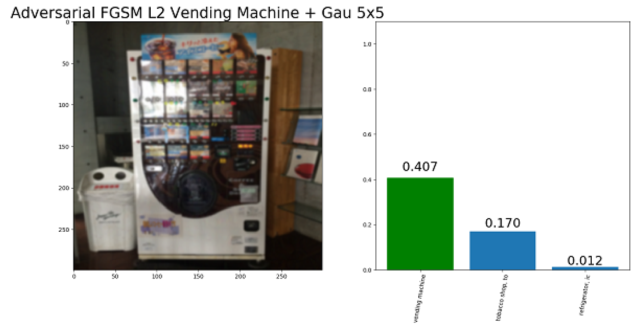
ing at scale," in *International Conference on Learning Representations ICLR*, 2017.

[34] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (S&P 2016)*. IEEE, 2016, pp. 582–597.

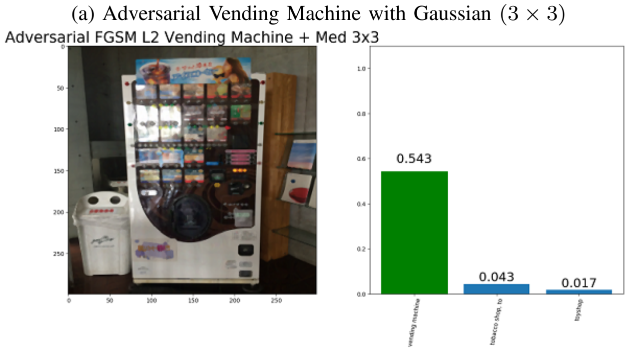
[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision IJCV*, vol. 115, no. 3, pp. 211–252, 2015.



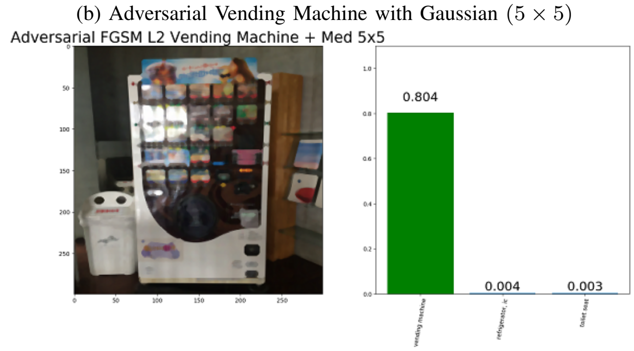
(a) Adversarial Vending Machine with Gaussian (3×3)



(b) Adversarial Vending Machine with Gaussian (5×5)

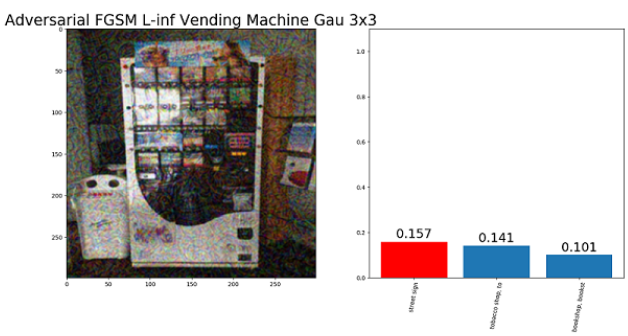


(c) Adversarial Vending Machine with Median (3×3)

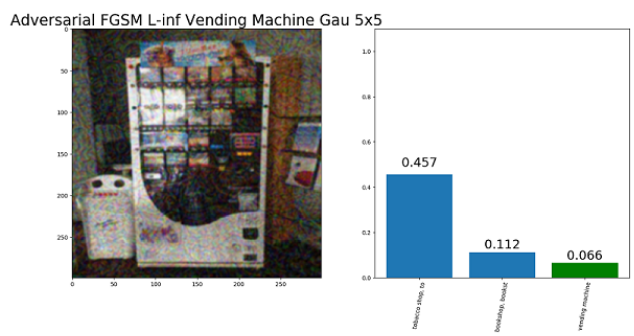


(d) Adversarial Vending Machine with Median (5×5)

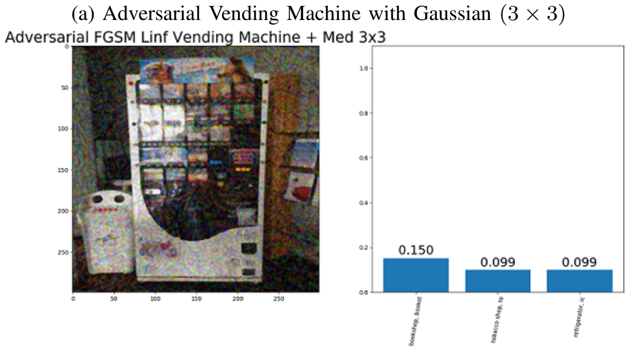
Fig. 7. Adversarial FGSM L_2 Vending Machine (targeted class: Street Sign) with Image Filters



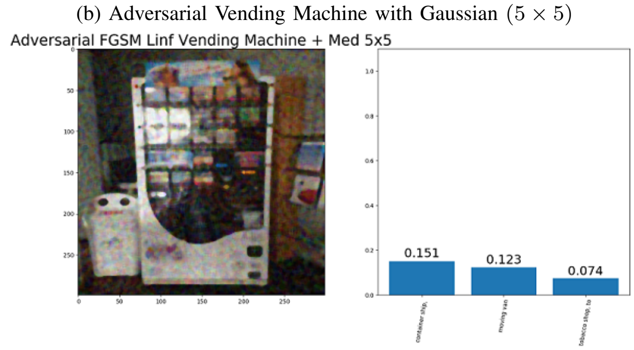
(a) Adversarial Vending Machine with Gaussian (3×3)



(b) Adversarial Vending Machine with Gaussian (5×5)



(c) Adversarial Vending Machine with Median (3×3)



(d) Adversarial Vending Machine with Median (5×5)

Fig. 8. Adversarial FGSM L_∞ Vending Machine (targeted class: Street Sign) with Image Filters

Towards Security Effectiveness Evaluation for Cloud Services Selection following a Risk-Driven Approach

Sarah Maroc[1], Jian Biao Zhang[2]
Beijing Key Laboratory of Trusted Computing
Faculty of Information Technology
Beijing University of Technology
Beijing, 100124, China

***Abstract*—Cloud computing is gaining a lot of popularity with an increasing number of services available in the market. This has rendered services selection and evaluation a difficult and challenging task, particularly for security-based evaluation. A key problem with much of the literature on cloud services security evaluation is that it fails to consider the overall evaluation context given the cloud characteristics and the underlying influence factors including threats, vulnerabilities, and security controls. In this paper, we propose a holistic risk-driven security evaluation approach for cloud services selection. We first use fuzzy DEMATEL method to jointly assess the likelihood and impact of threats with respect to the cloud service types, the exploitability of vulnerabilities to the identified threats, and the effectiveness of security controls in mitigating those vulnerabilities. Consequently, the overall diffusion of risk is captured via the relations across these concepts, which is leveraged to filter and prioritize the most critical security controls. The selected controls were then weighted using a combination of fuzzy DEMATEL and fuzzy ANP methods based on several factors, including their effectiveness in preventing the identified risks, user's preferences and level of control (i.e., responsibilities). The latter denotes how much control a cloud user is transferring to the cloud provider. To enhance the reliability of the results, the subjective weights were integrated with objective weights using the Entropy method. Finally, the TOPSIS method was employed for services ranking and the Improvement Gap Analysis (IGA) method was leveraged to provide more insights on the strength and weaknesses of the selected services. An illustrative example is given to demonstrate the application of the proposed framework.**

Keywords*—Cloud computing; cloud services selection; decision-making; risk-driven assessment; security evaluation

I. Introduction

Cloud computing has become increasingly popular due to its cost-effective and resources efficient services. It is a "model for enabling ubiquitous, convenient, and on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or cloud provider interaction" [1]. With the high number of cloud services available in the market, services selection and evaluation has become a significant challenge to users, particularly security evaluation.

The first and most critical step in any evaluation process is criteria identification. It describes the characteristics of the evaluation target that are of interest for the evaluation, thus it needs to be context specific. This is especially important in the cloud, given that the security threats, vulnerabilities, and controls differ from one service model to another. For example, IaaS suffers mainly from issues related to virtualization like hardening the host and securing inter-host communications. PaaS issues are more concerned with authentication and authorization. As for SaaS, the secure composition of the services is a critical area of concern [2]. Therefore, the selection of the critical security controls in the cloud needs to consider the overall dependencies between the vulnerabilities, threats, and the particular cloud services characteristics.

Various services selection methods have been proposed in the literature to support users in finding the most suitable services. However, in most of the available methods, the evaluation criteria were generally determined based on literature and experts' surveys (e.g., [3]–[8]). Other approaches have leveraged some recent standardization efforts such as the SMI framework [9] (e.g., [10], [11]), and CSA's CCM framework [12] (e.g., [13], [14]). Still, the evaluation criteria are generally specified in a rigid way for all cloud service models without considering the change in threats, and controls when applied to different cloud service models, rendering the evaluation process inefficient.

Indeed, an extensive list of criteria is important for a comprehensive evaluation. However, security evaluation is a challenging task that involves significant effort, in terms of both computational and human resources. Therefore, a minimal and representative set of evaluation criteria is more critical in a given context. This permits to focus on the situation and eliminate unnecessary tasks. Restricting the list of evaluation criteria will also help in the criteria weighting process. Most available weighting techniques such as AHP [15] and ANP [16] do not scale well with a large set of evaluation criteria, because of the large number of pairwise comparisons to be performed. For example, in the case of 20 evaluation criteria, 190 pairwise comparisons need to be performed using AHP method, which is both time consuming and a cumbersome task. The SMI framework [9] and CCM framework [12], which are widely used as evaluation criteria for cloud services selection, contain in total 51 attributes and 133 sub-controls, respectively. Thus, due to scalability issues, there is a need for a mechanism in place to first prioritize and select the most critical criteria given the evaluation context.

There exist several security standards on risk management such as NIST CRMF [17], which serve as good references for the selection of the baseline security controls. However, in these frameworks, the selection of the critical security controls is conducted in a purely qualitative way mostly relying on the expertise of the decision makers. With this challenge in mind, in this paper, we focus on selecting and prioritizing the critical security services in the cloud environment in a quantitative way following a risk management approach.

Following a risk-driven approach for cloud services selection helps in assessing the effectiveness of the security controls. Current cloud services evaluation and selection methods are mostly targeting the sufficiency and efficiency of the security controls, which focus on determining whether the security service performances meet customer's requirements. However, the presence of the security controls within the cloud service system does not necessarily mean it is always secure. Effectiveness measurement can only be appraised with sufficient knowledge about the threats and vulnerabilities [18]. Thus, adopting a risk-driven approach in selecting the evaluation criteria considering the relevant vulnerabilities and threats likelihoods, would consequently enable measurement of the effectiveness of the security controls. We further assess the extent of the effectiveness of the implemented security controls by analyzing their performance gaps against an assumed ideal using improvement gap analysis (IGA) method [19].

Another essential step related to the evaluation process is the weighting of the criteria. Current weighting approaches are generally based on the subjective users' preferences, criteria dependencies, or on the objective analysis of the evaluation data. An important factor that is not considered but highly relevant in the cloud context is the cloud users' varying degree of control over the implementation and management of the security services. In the cloud, the security responsibilities are shared among the cloud actors and depend on the cloud deployment model (i.e., public or private), service model (i.e., IaaS, PaaS, and SaaS), and the security control type. In the IaaS, the consumer is mostly responsible for securing the virtualized resources, application and data, while the cloud provider is responsible for securing the physical infrastructure. Contrary, in the SaaS, most of the security responsibilities are shifted to the provider side, leaving the consumer only responsible for the data and some minimal application management [17]. Accordingly, more importance should be assigned to the particular security control when the cloud user loses more control over its management to emphasis the responsibility for the associated security risks.

To summarize, the main contributions of this paper are:

1) Context-aware and risk-driven criteria selection. We benefit from our earlier work [20] on criteria selection for cloud services evaluation, with enhancements on the approach to address the scalability issues and account for the uncertainty and subjectivity of the process. In this paper, Fuzzy DEMATEL [21] method is used to identify the causal relationships between the cloud service types, threats and vulnerabilities, and the security controls. Fuzzy DEMATEL requires less comparisons compared to other dependency-aware techniques like ANP or AHP ($n(n-1)/2$). The goal is not to blindly use all the criteria that exist in the literature, but instead to identify those that are most critical to the context of the evaluation considering the characteristics of cloud service types and the overall evaluation context. This allows to reduce the effort required in the evaluation.

2) Criteria weighting considering more comprehensive set of factors. The proposed approach is distinguished from other existing methods in that it considers multiple factors in the weighting of criteria, namely user's preferences, criteria interdependencies, in addition to the user's level of control (i.e., responsibilities). The user's level of control reflects the degree of loss of control over the management and implementation of the security services, which represent one of the novelties of the proposed approach. Criteria weighting was performed using fuzzy DEMATEL and fuzzy ANP methods. The resultant subjective weights were further combined with objective weights based on Entropy method to obtain more accurate and less sensitive results to user's preferences or unreasonable criteria prioritization.

3) Effectiveness-driven evaluation following a risk-driven approach and gap analysis for performance improvement. The proposed framework attempts, on one hand, to enhance the efficiency of the evaluation process by reducing the set of evaluation criteria to the core attributes. On the other hand, it drives for an effectiveness-based evaluation of security services by assessing the effectiveness of the security controls in mitigating the potential threats and vulnerabilities, thus the risks prevented. Furthermore, the extent of the effectiveness of the implemented security controls are assessed using the improvement gap analysis (IGA) method [19].

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 presents the proposed framework. Section 4 demonstrates the effectiveness of the proposed approach through a case study. Section 5 concludes the paper.

II. Related Work

In this section, we will review the related work on cloud services evaluation with the focus on security-driven studies and dependency-aware cloud services selection approaches.

A. Cloud Services Security-based Evaluation

Security evaluation, in our context, aims to provide a quantification of the security level of cloud services in a way to enable comparison between different services offerings. Cloud services evaluation has mostly targeted measurable attributes such as performance and availability, with less focus on security [22], [23]. Although security is mentioned in almost every study on cloud services evaluation, most of the studies do not of focus on security related attributes and influence factors.

Among the few works focused on security, Mouratidis et al. [24] proposed a holistic framework starting from the elicitation of the security and privacy requirements to the selection of cloud services providers. Luna et al. [13] presented two evaluation techniques, namely Quantitative Policy Trees (QPT) and Quantitative Hierarchy Tree (QHP) for assessing the security level of cloud providers as per the claimed SLAs with respect to users' requirements. The QPT weighted the criteria and aggregated alternatives performances in an ad-hoc manner, whereas QHP employed the AHP technique for criteria weighting and ranking of alternatives. Modic et al. [14] proposed a cloud security assessment technique called Moving Intervals Process (MIP) that aimed at decreasing the time complexity of the assessment algorithm by separating scores for services providers that can fulfill users' needs from scores of those that are under-provisioning. Halabi and Bellaiche [8] proposed a security self-evaluation methodology for cloud providers using a variety of security metrics. In another work from the same authors [3], the security level of cloud service providers was quantified with respect to the traditional security attributes (CIA triad), namely: confidentiality, integrity, and availability. The best solution was then obtained using a linear multi-objective optimization technique that aims at minimizing the dissatisfaction factors. Alabool and Mahmood [25] proposed a framework for ranking IaaS cloud providers and used the IPA method for ranking the unimproved gaps to provide insights on how to better improve the cloud services.

In the above studies and most available cloud services evaluation approaches, an extensive list of criteria is employed in the evaluation, either identified through literature and experts survey, or by leveraging existing frameworks such as the SMI framework [9] or CCM framework [12]. However, these frameworks target cloud services in general and do not consider the change in threats and measures when applied to different cloud deployment models and service types. Besides, the long list of criteria (e.g., CCM framework includes 16 control domains with more than 130 security sub-controls) renders the weighting process a tedious task. Furthermore, security evaluation constitutes only a part of the overall trustworthiness evaluation of cloud services. A variety of other evaluation criteria including financial and performance attributes are of interest. Thus, prioritizing and filtering the criteria to a minimum and representative set is important for practical and efficient evaluation.

There exists some general security frameworks and guidelines for selecting baseline security controls such as NIST cloud-adapted risk management framework (CRMF) [17]. However, existing standards lack a quantitative and systematic method of how controls should be selected. In [26], the authors proposed a quantitative framework for prioritizing the security controls with respect to the identified vulnerabilities and threats given the severity and cost of the remediation effort. Nevertheless, the proposed framework targets the security information domain in general and thus fails to consider the specific characteristic of the cloud environment including the influences of cloud service models on the potential threats and vulnerabilities, as well as the shared responsibility of cloud users in the process. In the next section, we will discuss some of the works addressing dependency relations in cloud security evaluation literature.

B. Dependency-Aware Cloud Security Evaluation Methods

The relationships between the evaluation concepts are often neglected in existing cloud evaluation studies. To address this lack, Sun et al. [27] applied fuzzy measure and Choquet integral to measure and aggregate non-linear relations between criteria. Taha et al. [28] proposed a framework for measuring the structural dependencies between cloud security services, which were then used as weights for the evaluation criteria. However, the proposed approach only considered the relations between the services in a hierarchical structure. In [29], The influences of attributes on the overall quality of services were integrated with the user's preferences in order to calculate the final weights of attributes using the ANP method to allow for a flexible network-like structure representation. In [30], the authors employed fuzzy-ANP to calculate criteria weights for cloud services evaluation. In [31], the authors examined the causal relationships between the criteria using fuzzy DEMATEL-based ANP technique to determine the influence and the weights of the criteria. VIKOR method was then employed to rank the alternatives and identify the weaknesses to help improve service performances. Several other works have combined DEMATEL and ANP to handle the dependencies between the evaluation criteria in the cloud such as [32] and [33]. However, the above-reviewed methods do not consider the dependencies between criteria from a risk perspective.

In [34], the authors applied DEMATEL-based ANP to account for the dependencies between the security controls, which were identified following risk assessment procedure. Also, in [35], a method was proposed for evaluating the risk levels of information security. DEMATEL was first used to analyze the interrelations among security control areas. The risk likelihood ratings were then obtained using the ANP method. Still, these frameworks only considered the dependencies between the security controls directly. That is, the influence of threats and vulnerabilities were not jointly included in the quantitative analysis. Besides, the above methods were applied to security information in general, and hence lack the specific characteristic of the cloud environment. That is the change in threats, vulnerabilities, and controls when applied to different cloud deployment and service models types.

Overall, while some researchers have considered the dependencies between the evaluation criteria, they have ignored the characteristics of cloud service model types, as well as the underlying risk factors (threat likelihood, vulnerability relevance, and control effectiveness). Besides, the dependencies between criteria, when considered, were only addressed as part of the weighting process. In contrast, in this paper, we leverage the causal relationships between the cloud service types, threats vulnerabilities and security controls to extract the minimum and critical set of the evaluation criteria. The dependency values were then integrated with users' preferences and their level of control (i.e., responsibilities) to obtain the total subjective weights, which were then combined with objective weights to improve the reliability and accuracy of the approach. The proposed framework attempts to enhance the efficiency of the evaluation process by reducing the set of evaluation criteria to the core factors, and drive for effectiveness-based evaluation by understanding the extent of the effectiveness of the implemented security controls in preventing the risks.

III. Proposed Framework

The proposed framework, as shown in Fig. 1, consists of five main phases: context establishment, criteria selection, criteria weighting, services ranking, and finally performance improvement and gap analysis. The detailed description of the steps at each phase is described in the following sections.

A. Context Building and Criteria Selection

The concepts model follows a risk perspective by modeling the threats, vulnerabilities, and security controls, while considering the characteristics of the cloud service types. The problem can be formally modeled as follows. Let $S = \{s_1, s_2, \dots, s_w\}$ be the cloud service model types of the evaluation target, $T = \{t_1, t_2, \dots, t_p\}$ the threats, $V = \{v_1, v_2, \dots, v_q\}$ the vulnerabilities, and $C = \{c_1, c_2, \dots, c_n\}$ the security controls representing the evaluation criteria.

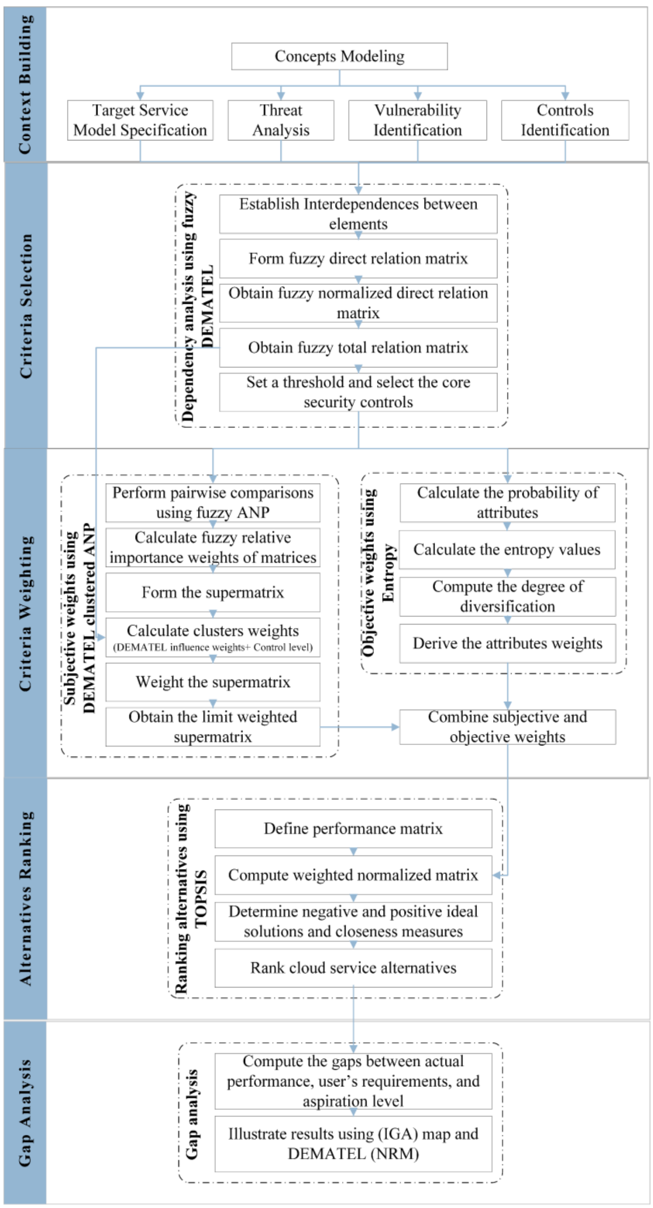


Fig. 1. The Conceptual Model of the Proposed Framework.

For criteria selection, DEMATEL [21] method is used to analyze the dependencies (direct and indirect) between the service model, potential threats, exploited vulnerabilities, and the appropriate security controls. This way, we can jointly assess the likelihood of threats given the service type, the relevance of various vulnerabilities to the identified threats, and the effectiveness of the security controls in mitigating the vulnerabilities. Consequently, the overall diffusion of risk is captured via the relations and dependencies across these concepts, which will be used to filter and prioritize the critical security controls that contribute the most to the evaluation. To cope with the fact that human judgment is often uncertain and hard to estimate by exact numerical values. fuzzy theory [36] is applied to the DEMATEL method. The output at this stage is a list of the minimal and critical security controls judged necessary and sufficient for an effective and efficient evaluation of cloud services. The steps are as follows.

Step 1. Establishing the dependencies between elements and forming the fuzzy direct-relation matrix. The direct-relation matrix $\tilde{Z}$ is constructed through pairwise comparison among the elements in which $\tilde{z}_{ij} = (l_{ij}, m_{ij}, u_{ij})$ indicates the degree to which the element i affects element j as ascertain by experts. It is assumed that a consensus of opinions exists among experts in the evaluation process.

$$\tilde{Z} = \begin{array}{cc} & \begin{array}{ccccccccccccc} & S & & & T & & & V & & & C & \\ s_1 & \cdots & s_w & t_1 & \cdots & t_p & v_1 & \cdots & v_q & c_1 & \cdots & c_n \end{array} \\ \begin{array}{cc} S & \begin{array}{c} s_1 \\ \vdots \\ s_w \end{array} \\ T & \begin{array}{c} t_1 \\ \vdots \\ t_p \end{array} \\ V & \begin{array}{c} v_1 \\ \vdots \\ v_q \end{array} \\ C & \begin{array}{c} c_1 \\ \vdots \\ c_n \end{array} \end{array} & \begin{bmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \tilde{z}_{ij} \\ \vdots & & & & & & & & & & & \cdot \\ \tilde{z}_{ij} & & & & & & & & & & & \cdot \\ \cdot & & & \cdot & & & & & & & & \cdot \\ \cdot & & & & \cdot & & & & & & & \cdot \\ \cdot & & & & & \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot & & & & & \cdot \\ \cdot & & & & & & & \cdot & & & & \cdot \\ \cdot & & & & & & & & \cdot & & & \cdot \\ \cdot & & & & & & & & & & & \cdot \\ \cdot & & & & & & & & & & & \cdot \\ \tilde{z}_{ij} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix} \end{array} \quad (1)$$

Fig. 2 illustrates the graphical structure of the concepts and their relations.

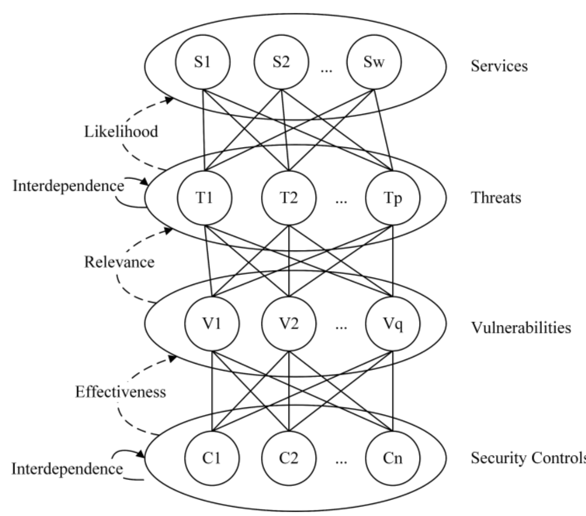


Fig. 2. Criteria Selection Problem Structure of the underlying Factors and Relations.

TABLE. I. INFLUENCE MEASURES USING LINGUISTIC TERMS.

Linguistic terms	Linguistic values
No influence (N)	(0, 0, 0.25)
Low influence (L)	(0, 0.25, 0.5)
Medium influence (M)	(0.25, 0.5, 0.75)
High influence (H)	(0.5, 0.75, 1)
Very high influence (VH)	(0.75, 1, 1)

The fuzzy linguistic scale is used for measuring the influence degree with its equivalent triangular membership function, as shown for example in Table I.

Step 2. Calculating the normalized fuzzy direct-relation matrix. On the base of the direct-relation matrix $\tilde{Z}$, the normalized direct-relation matrix $\tilde{X}$ can be obtained as follows.

$$\tilde{z}_{ij} = \left(l_{ij}, m_{ij}, u_{ij}\right) \text{ and } s = \frac{1}{\max\limits_{1 \le i \le m} \sum_{j=1}^{m} u_{ij}}, \text{ then} \tag{2}$$

$\tilde{X} = \tilde{Z} \times s$, where

$$\tilde{X} = [\tilde{x}_{ij}]_{m \times m} = \tilde{z}_{ij} \times s = (l_{ij} \times s, m_{ij} \times s, u_{ij} \times s)$$

Step 3. Calculating the fuzzy total-relation matrix. The fuzzy direct/indirect relation matrix, known as the total relation matrix can be obtained as follows.

$$\tilde{T} = \lim_{k \to \infty} (\tilde{X} + \tilde{X}^2 + \cdots + \tilde{X}^k) = \tilde{X}\left(I - \tilde{X}\right)^{-1} \tag{3}$$

$\tilde{T} = \left[\tilde{t}_{ij}\right]_{m \times m}$, $\tilde{t}_{ij} = \left(\grave{l}_{ij}, \grave{m}_{ij}, \grave{u}_{ij}\right)$, where

$[\grave{l}_{ij}]= \tilde{X}_l\left(I - \tilde{X}_l\right)^{-1}$, $[\grave{m}_{ij}]= \tilde{X}_m\left(I - \tilde{X}_m\right)^{-1}$, $[\grave{u}_{ij}]= \tilde{X}_u\left(I - \tilde{X}_u\right)^{-1}$

Where I is the identity matrix.

Step 4. Setting a threshold value and selecting the critical security controls. A threshold value α is set to filter minor effects and reduce the complexity of the decision process. Only elements whose influence value in the total matrix is higher than the threshold value can be chosen. The influence values in matrix $\tilde{T}$ is reset to zero if its values are less than α. The new matrix is called the α-cut total-influence matrix $\tilde{T}_{\alpha}$. Based on this idea, we exclude the security controls with negligible effects and select the controls with the most influence relationships. The threshold value can be decided by experts or using analytical methods such as the mean value of the total influence matrix. To simplify the calculation, we first defuzzify the total fuzzy relation matrix $\tilde{T}$. Several defuzzification methods exist, we chose the center of area (CoA) method, as it is the most commonly used method. The formula is as follows.

$$t_{ij} = \frac{(\grave{u}_{ij} - \grave{l}_{ij} + \grave{m}_{ij} - \grave{l}_{ij})}{3} + \grave{l}_{ij} \tag{4}$$

The sum of rows R_i denotes the sum of direct and indirect effects of element i on the other elements. Whereas, the sum of columns D_j denotes the sum of direct and indirect effects that element j has received from the other elements. Consequently, $R_i + D_i$ denotes the strength of influences given and received, which represents the degree of the central role that element i plays in the decision-making process. If $R_i - D_i$ is positive than element i is affecting other elements (cause group), if negative, it is being influenced by the other elements (effect group). Furthermore, a visual causal diagram can be depicted by arranging $R_i + D_i$ values in x-axis and $R_i - D_i$ values on the y-axis.

B. Criteria Weighting

The selected security controls from the previous stage represent the top-level evaluation criteria (dimensions or clusters). These criteria are further divided into more fine-grained sub-criteria. The weights of criteria are calculated using subjective and objective methods. The subjective weights are determined based on the influence degree of the criteria, level of control, and their importance to the users. Fuzzy ANP method is used to assign the importance weights to the criteria through pairwise comparisons. However, contrary to the assumption of equal cluster's weight in traditional ANP, we use fuzzy DEMATEL influence degrees obtained previously combined with the level of control degree to weight the clusters. The obtained subjective weighs are further adjusted with objective weights using the entropy method to obtain more reliable results. The steps for weighting the criteria are described below.

Step 1. Performing pairwise comparison and obtaining priority vectors. The ANP method [16] combined with fuzzy set theory is employed to derive the subjective weights. The relations between clusters (i.e., dependence relations between security controls) are determined based on the previous results from the DEMATEL network relation map (NRM). Once the relations between criteria and sub-criteria are identified, users are asked to perform pairwise comparison between criteria. The importance values are assigned using triangular fuzzy numbers based on a 9-point scale (from equally important to extremely important) The priority vectors for each pairwise comparison matrix can be calculated using the eigenvalue method [16]. Then, the weighs are defuzzified in the same way as in Eq. (4). A consistency ratio of the pair-wise comparisons is calculated and should be less than 0.10 for the comparison to be acceptable. Otherwise, it is necessary to adjust the results. The priorities are gathered into the appropriate columns to build the supermatrix. The form of the supermatrix is as follows.

$$W = \begin{array}{cc|c} & & \begin{array}{ccc} C_1 & C_j & C_n \\ c_{11} \cdots c_{1m_1} & c_{1j} \cdots c_{jm_j} & c_{1n} \cdots c_{nm_n} \end{array} \\ C_1 & \begin{array}{c} c_{11} \\ \vdots \\ c_{1m_1} \end{array} & \\ C_i & \begin{array}{c} c_{i1} \\ \vdots \\ c_{im_1} \end{array} & \begin{bmatrix} W_{11} & \dots & W_{1j} & \dots & W_{1n} \\ \vdots & & \vdots & & \vdots \\ W_{i1} & \dots & W_{ij} & \dots & W_{in} \\ \vdots & & \vdots & & \vdots \\ W_{n1} & \dots & W_{nj} & \dots & W_{nn} \end{bmatrix} \\ C_n & \begin{array}{c} c_{n1} \\ \vdots \\ c_{nm_n} \end{array} & \end{array} \tag{5}$$

Step 2. Obtaining the weights of clusters. In the traditional ANP, the weights of elements are divided by the number of clusters. This normalization method implies that the clusters are of equal weights (in our context the high-level security controls). However, in reality, the effect of each cluster on the other clusters is different, and have been determined in the previous step using fuzzy DEMATEL method. Hence, these influence values are used in weighting the clusters, in addition to another factor, which is the level of control.

Step 2.1. Obtaining the influence degree of the clusters. The interdependencies between the clusters are already determined previously using DEMATEL, hence can be directly derived from the total influence matrix. Let $T^{\alpha SC}$ be the α-cut total-influence matrix for security controls. $t_{ij}^{\alpha SC}$ represents the degree of influence that the cluster i (i.e., security control) exerts on the cluster j.

Step 2.2. Determining the user's level of control degree. The degree of control (w^{cont}) denotes how much control a consumer is transferring to the cloud provider. Accordingly, more importance should be assigned to the particular security control when the cloud user loses more control over its management, as oppose to when the user has full control for its management. In NIST security reference architecture [17], the responsibility of the cloud user for each security component given the cloud deployment model and service type was defined as follows:

- Full responsibility. Meaning the user has full control over the management of the security control and thus, less importance value should be assigned to the security control. In this case ($w^{cont} = 0.25$).
- Shared responsibility. Meaning both the cloud user and provider share responsibility for managing the particular security control. In this case ($w^{cont} = 0.5$).
- Least responsibility. Meaning the provider has full control over the management of the security control. The consumer needs to negotiate with the provider to ensure that the requirements are met. Therefore, more importance value is assigned to the security control since the consumer loses the ability to implement it and manage it. In this case ($w^{cont} = 1$).

For example, the responsibility for the security component "Data Governance >Secure Disposal of Data" is a shared responsibility between the consumer and provider in the IaaS model ($w^{cont} = 0.5$), but needs to be implemented by the cloud provider in all other service models ($w^{cont} = 1$).

Step 2.3. Obtaining the total weights of the clusters. The weight of the cluster w^{cl} is the product of its influence degree $T^{\alpha SC}$ and level of control w^{cont}.

$$w_{ij}^{cl} = w_i^{cont} \times t_{ij}^{\alpha SC} \tag{6}$$

The clusters' weights are then normalized as follows.

$$w_{ij}^{cl} = \frac{w_{ij}^{cl}}{\sum_{j=1}^{n} w_{ij}^{cl}} \tag{7}$$

Step 3. Obtaining the weighted supermatrix. By combining the weights of the clusters with the unweighted supermatrix as defined in [37], we obtain the weighted supermatrix as follows.

$$W^s = w^{cl}W = \begin{array}{cc} & \begin{array}{ccc} C_1 & C_j & C_n \\ c_{11}\cdots c_{1m_1} & c_{1j}\cdots c_{jm_j} & c_{1n}\cdots c_{nm_n} \end{array} \\ \begin{array}{c} C_1 \begin{array}{c} c_{11} \\ \vdots \\ c_{1m_1} \end{array} \\ C_i \begin{array}{c} c_{i1} \\ \vdots \\ c_{im_1} \end{array} \\ C_n \begin{array}{c} c_{n1} \\ \vdots \\ c_{nm_n} \end{array} \end{array} & \begin{bmatrix} w_{11}^{cl} \times W_{11} & \dots & w_{j1}^{cl} \times W_{1j} & \dots & w_{n1}^{cl} \times W_{1n} \\ \vdots & & \vdots & & \vdots \\ w_{1i}^{cl} \times W_{i1} & \dots & w_{ji}^{cl} \times W_{ij} & \dots & w_{ni}^{cl} \times W_{in} \\ \vdots & & \vdots & & \vdots \\ w_{1n}^{cl} \times W_{n1} & \dots & w_{jn}^{cl} \times W_{nj} & \dots & w_{nn}^{cl} \times W_{nn} \end{bmatrix} \end{array} \tag{8}$$

Step 4. Obtaining the limit weighted supermatrix. To obtain the global priorities, the weighted supermatrix is raised to the limiting powers $\lim_{k\to\infty} W_s^k$, where k is the number of powers.

Step 5. Calculating the objective weights. In the previous steps, criteria weights were calculated using subjective approaches and based on subjective factors that rely heavenly on decision-makers' opinions. To adjust the weights and help achieve more reliable results, we measure the weights using objective method, namely, the entropy method [38]. The entropy method determines the criterion's weight based on the information transmitted by that criterion. That is, if a particular criterion has similar values for all the alternatives, then this criterion has little importance in the decision-making. In contrast, the criterion that alternatives are most dissimilar should have the highest importance weight since it transmits more information and helps to differentiate between the different alternatives.

The projected outcomes P_{ij} of a criterion c_j is defined as:

$$P_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \tag{9}$$

x_{ij} is the performance of alternative i on criterion j.

The entropy is calculated as follows:

$$ET_j = -\left(\frac{1}{\ln m}\right) \sum_{i=1}^{m} P_{ij} \ln P_{ij} \tag{10}$$

The degree of diversification of the information provided by the criterion j is

$$d_j = 1 - ET_j \tag{11}$$

The entropy weight is then:

$$w_j^o = \frac{d_j}{\sum_{j=1}^{n} dj} \tag{12}$$

Step 6. Compute the final criteria weights. The final criteria weights are obtained by combing the subjective and objective weights as follows.

$$w_j = \alpha w_j{}^o + \beta w_j{}^s, \text{ where } \alpha + \beta = 1 \tag{13}$$

$\alpha \; and \; \beta$ can be adjusted accordingly to reflect the influence of subjective and objective weights on the decision-making.

C. Services Ranking

After weighing the criteria, the ranking of the best cloud service provider is performed using TOPSIS [39] method. TOPSIS method is based on the distance measure of an alternative from the ideal solution, taking into account both the closeness distance from the positive ideal solution (PIS) and the farthest distance from the negative ideal solution (NIS). TOPSIS was chosen as it best reflects the risk attitudes of decision-makers. The smaller the distance measure from PIS, the higher alternative preference to profit; whereas the larger the distance measure from NIS, the higher the alternative preference to avoid risk [39]. This approach is suitable for a security-based evaluation of cloud services as a risk avoider strategy. Due to space limitations, the steps of TOPSIS method can be found in [39].

D. Performance Improvement and Gap Analysis

Most existing cloud services evaluation studies have limited the evaluation process to the ranking of cloud services alternatives. However, the evaluation process also aims to help cloud service providers in improving their service performances. Few studies have attempted to identify what should be improved. Work in this direction was proposed by Alabool et al. [4]. They used the importance-performance analysis (IPA) [40] method to identify and rank the unimproved gaps. IPA is one of the most used methods to identify the strength and weaknesses of service performances. However, IPA has some limitations concerning the nonlinearity between the performance of attributes and customer satisfaction [19]. Aiming to overcome these problems, Tontini and Picolo [19] proposed the improvement gap analysis (IGA) method. IPA method compares the performance of the criteria with respect to their importance. In contrast, the IGA method compares the expected customer dissatisfaction if an attribute has a low performance with the expected customer satisfaction if the attribute is improved [19].

In traditional IGA, customers are asked to estimate their expected satisfaction and dissatisfaction with respect to each attribute and the actual attribute performance. The improvement gap (IG) for each attribute is calculated as the difference between the expected and the actual performance ($IG = EP - AP$) . The dissatisfaction is stated directly according to the expected impact on customer dissatisfaction if an attribute has low performance. In this paper, we calculate the improvement gap as the difference between the best available performance among all alternatives (BP_j) and the actual performance of the particular service AP_{ij} (Eq. 14). The value of the gap represents the scope of improvement needed in order to achieve high market competition. The best performance can also be replaced by the aspirational levels instead of the minimum-maximum values.

$$IG_{ij} = BP_j - AP_{ij} \text{ where} \tag{14}$$

$$BP_j = (\max(y_{ij}) \mid j \in Benefit, \min(y_{ij}) \mid j \in Cost))$$

y_{ij} represents the performance of alternative i on criteria j.

As for the dissatisfaction value (DP_{ij}), it is calculated based on its importance and the difference between user's requirements (RQ_j) and the actual service performance.

$$DP_{ij} = w_j * (RQ_j - AP_{ij}) \tag{15}$$

We plot the performance of alternatives into a two-dimensional graph as defined in the IGA method (see Fig. 3), showing each criterion's expected dissatisfaction on the y-axis with respect to the improvement gap on the x-axis. Attributes are classified into four categories: (1) critical for improvement, (2) keep as it is, (3) attractive, and (4), neutral [19].

An attribute is classified as critical to improve if its performance is lower than its competitors and doesn't satisfy the customer's requirements. It is classified as keep as it is when its performance is higher than the competition but not fully satisfying customers' requirements. Employing more resources to improve this attribute when its performance is already higher and deemed sufficient than the market, will not necessarily bring superior satisfaction to costumers, which may lead to a waste of resources. It is classified as attractive attribute if there is no strong dissatisfaction with its performance but there is still a high gap to the market, which if improved can bring superior customer satisfaction. It is classified as neutral when more improvement in this attribute will neither bring strong market differentiation nor superior customer satisfaction.

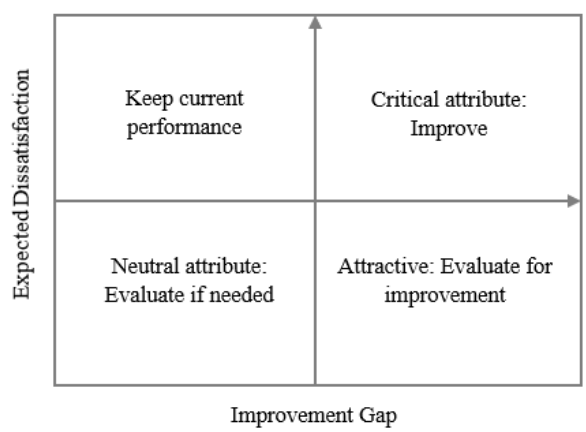


Fig. 3. Adapted Importance-Performance Analysis (IGA) Map [19].

IV. Case Study

To demonstrate the applicability of the proposed framework, we present an example of an evaluation for a SaaS service using data extracted from NIST security reference architecture [17] and CSA STAR repository, which is a public registry that documents the security controls provided by popular cloud computing. Following the proposed framework, the first phase involves establishment of the evaluation context, including the modeling of the target's service model, potential threats, vulnerabilities, and available security controls. For simplicity and without loss of generality, we consider the list of possible threats, vulnerabilities, and security controls presented in Table II. The next phase involves the selection of the critical security controls from the derived list of controls.

TABLE. II. EVALUATION CONCEPTS

(List of threats, vulnerabilities, and controls)	
	Designation
T1	Denial of service
T2	Data leakage
T3	Account or service hijacking
T4	Malicious insiders
T5	Cross VM attacks
T6	Sniffing /spoofing virtual networks
T7	Insecure VM migration
V1	Unlimited allocation of resources
V2	Incomplete data removal
V3	Authentication & authorization vulnerabilities
V4	VM co-residence
V5	Data collocation with weak separation
V6	Insecure interfaces and APIs
V7	Communication encryption vulnerabilities
C1	Application security
C2	Data security
C3	Identity & access management
C4	Human resources
C5	Virtualization security
C6	Security monitoring services
C7	Information system regulatory mapping

The initial fuzzy direct relation matrix of DEMATEL is shown in Table III using linguistic values from Table I. It depicts the different dependencies between the cloud service type, threats, vulnerabilities, and security controls considering several factors: the likelihood of a threat on SaaS service type, its impact, the relevance degree of the vulnerabilities to the identified threats, the effectiveness of controls on mitigating those vulnerabilities, and the interdependencies between the security controls. Following steps 2-4 (Section 3.1), we obtain the defuzzified total influence matrix, as shown in Table IV. The resultant security controls submatrix is depicted in bold in Table IV. We set a threshold value of (0.068); influence values less than the threshold are reset to zero. From the results, it can be concluded that criterion (C4) have less impact and relevance on the overall evaluation in this case study, thus it is excluded from the evaluation process. The resulting network relation (NRM) structure between the selected security controls (C1, C2, C3, C5, C6, and C7) is shown in Fig. 4.

After the selection of the critical security controls and establishing the network structure, we proceed to the next phase, criteria weighting. The control criteria are divided into more fine-grained sub-criteria. Table V presents the performance of alternatives with respect to the criteria. After performing the pairwise comparisons between each node in the cluster and the nodes in the related clusters as per the network structure, we obtain the initial supermatrix (step 1, section 3.2), as shown in Table VI.

TABLE. III. THE INITIAL FUZZY RELATION MATRIX

	SaaS	*T1*	*T2*	*T3*	*T4*	*T5*	*T6*	*T7*	*V1*	*V2*	*V3*	*V4*	*V5*	*V6*	*V7*	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*
SaaS	0	M	M	L	M	M	M	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T1	H	0	M		L	M	M	L	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T2	H	L	0	M	L	M	M	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T3	H	M	H	0	L	M	M	L	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T4	H	L	H	H	0	M	M	L	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T5	H	M	H	H	L	0	M	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T6	H	M	H	M	L	M	0	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T7	H	L	H	M	L	M	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V1	0	H	M	M	L	L	L	L	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V2	0	M	M	L	L	L	L	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V3	0	H	H	H	H	H	H	H	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V4	0	M	H	H	L	H	M	H	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V5	0	L	H	H	L	H	M	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V6	0	H	M	M	H	M	H	H	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V7	0	M	M	M	H	M	H	H	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C1	0	0	0	0	0	0	0	0	H	H	M	M	H	H	M	0	L	H	L	M	M	L
C2	0	0	0	0	0	0	0	0	L	M	H	H	H	H	H	H	0	H	H	H	L	M
C3	0	0	0	0	0	0	0	0	M	H	M	M	M	H	M	H	H	0	H	H	L	M
C4	0	0	0	0	0	0	0	0	L	L	L	M	L	L	L	L	M	M	0	L	L	M
C5	0	0	0	0	0	0	0	0	L	H	H	H	H	M	H	H	H	M	M	0	M	M
C6	0	0	0	0	0	0	0	0	L	L	M	H	M	M	M	M	L	L	M	M	0	H
C7	0	0	0	0	0	0	0	0	L	M	M	M	M	M	M	L	L	M	H	M	M	0

TABLE. IV. THE DEFUZZIFIED TOTAL RELATION MATRIX

	SaaS	***T1***	***T2***	***T3***	***T4***	***T5***	***T6***	***T7***	***V1***	***V2***	***V3***	***V4***	***V5***	***V6***	***V7***	***C1***	***C2***	***C3***	***C4***	***C5***	***C6***	***C7***
SaaS	0.0393	0.0696	0.0799	0.0573	0.0646	0.0731	0.0730	0.0706	0.0225	0.0261	0.0261	0.0270	0.0265	0.0265	0.0261	0.0252	0.0244	0.0252	0.0256	0.0252	0.0234	0.0247
T1	0.0917	0.0346	0.0795	0.0739	0.0474	0.0730	0.0730	0.0533	0.0225	0.0261	0.0261	0.0270	0.0265	0.0265	0.0261	0.0252	0.0243	0.0252	0.0256	0.0252	0.0234	0.0247
T2	0.0917	0.0526	0.0444	0.0739	0.0474	0.0730	0.0730	0.0709	0.0225	0.0261	0.0261	0.0270	0.0265	0.0265	0.0261	0.0252	0.0243	0.0252	0.0256	0.0252	0.0234	0.0247
T3	0.0933	0.0711	0.0982	0.0397	0.0482	0.0743	0.0743	0.0546	0.0229	0.0266	0.0265	0.0274	0.0270	0.0270	0.0265	0.0256	0.0248	0.0256	0.0260	0.0256	0.0238	0.0251
T4	0.0950	0.0548	0.1003	0.0941	0.0311	0.0757	0.0757	0.0556	0.0233	0.0271	0.0270	0.0279	0.0275	0.0275	0.0270	0.0261	0.0252	0.0261	0.0265	0.0261	0.0243	0.0256
T5	0.0966	0.0733	0.1017	0.0951	0.0499	0.0414	0.0769	0.0740	0.0237	0.0275	0.0275	0.0284	0.0279	0.0279	0.0275	0.0265	0.0256	0.0265	0.0270	0.0265	0.0247	0.0260
T6	0.0949	0.0720	0.0999	0.0765	0.0490	0.0756	0.0401	0.0731	0.0233	0.0270	0.0270	0.0279	0.0275	0.0275	0.0270	0.0261	0.0252	0.0261	0.0265	0.0261	0.0243	0.0256
T7	0.0933	0.0535	0.0985	0.0752	0.0482	0.0743	0.0743	0.0366	0.0229	0.0266	0.0265	0.0274	0.0270	0.0270	0.0265	0.0256	0.0248	0.0256	0.0261	0.0256	0.0238	0.0252
V1	0.0418	0.0853	0.0758	0.0716	0.0447	0.0525	0.0525	0.0498	0.0156	0.0251	0.0251	0.0259	0.0255	0.0255	0.0251	0.0242	0.0234	0.0242	0.0246	0.0242	0.0225	0.0238
V2	0.0401	0.0659	0.0744	0.0525	0.0439	0.0511	0.0511	0.0669	0.0212	0.0186	0.0246	0.0254	0.0250	0.0250	0.0246	0.0238	0.0230	0.0238	0.0242	0.0237	0.0221	0.0233
V3	0.0591	0.0968	0.1115	0.1045	0.0898	0.1017	0.1017	0.0976	0.0258	0.0300	0.0239	0.0310	0.0305	0.0305	0.0300	0.0289	0.0280	0.0290	0.0294	0.0289	0.0269	0.0284
V4	0.0522	0.0744	0.1045	0.0983	0.0501	0.0963	0.0785	0.0933	0.0242	0.0281	0.0280	0.0229	0.0285	0.0285	0.0280	0.0270	0.0262	0.0271	0.0275	0.0270	0.0252	0.0266
V5	0.0488	0.0544	0.1013	0.0956	0.0483	0.0936	0.0758	0.0733	0.0233	0.0271	0.0271	0.0280	0.0215	0.0275	0.0271	0.0261	0.0253	0.0262	0.0266	0.0261	0.0243	0.0256
V6	0.0539	0.0932	0.0886	0.0823	0.0871	0.0799	0.0976	0.0940	0.0246	0.0286	0.0285	0.0295	0.0290	0.0230	0.0285	0.0275	0.0266	0.0276	0.0280	0.0275	0.0256	0.0270
V7	0.0522	0.0742	0.0872	0.0810	0.0863	0.0785	0.0963	0.0930	0.0242	0.0281	0.0280	0.0290	0.0285	0.0285	0.0220	0.0271	0.0262	0.0271	0.0275	0.0271	0.0252	0.0266
C1	0.0473	0.0560	0.0665	0.0609	0.0500	0.0593	0.0592	0.0592	0.0328	0.0896	0.0711	0.0733	0.0902	0.0905	0.0711	**0.0339**	**0.0498**	**0.0858**	**0.0524**	**0.0690**	**0.0642**	**0.0502**
C2	0.0543	0.0654	0.0775	0.0715	0.0583	0.0693	0.0693	0.0690	0.0496	0.0801	0.0968	0.0995	0.0985	0.0985	0.0968	**0.0939**	**0.0390**	**0.0943**	**0.0949**	**0.0936**	**0.0532**	**0.0743**
C3	0.0520	0.0628	0.0731	0.0670	0.0551	0.0647	0.0649	0.0648	0.0661	0.0965	0.0782	0.0809	0.0799	0.0974	0.0782	**0.0929**	**0.0904**	**0.0408**	**0.0939**	**0.0925**	**0.0522**	**0.0733**
C4	0.0400	0.0455	0.0534	0.0493	0.0402	0.0477	0.0474	0.0471	0.0406	0.0483	0.0482	0.0679	0.0491	0.0494	0.0482	**0.0465**	**0.0623**	**0.0642**	**0.0300**	**0.0468**	**0.0424**	**0.0629**
C5	0.0533	0.0639	0.0762	0.0700	0.0568	0.0679	0.0675	0.0676	0.0484	0.0959	0.0957	0.0984	0.0974	0.0797	0.0957	**0.0926**	**0.0895**	**0.0758**	**0.0764**	**0.0399**	**0.0695**	**0.0734**
C6	0.0462	0.0542	0.0642	0.0594	0.0482	0.0576	0.0572	0.0569	0.0437	0.0532	0.0708	0.0910	0.0720	0.0717	0.0708	**0.0678**	**0.0486**	**0.0513**	**0.0697**	**0.0681**	**0.0291**	**0.0848**
C7	0.0459	0.0540	0.0636	0.0585	0.0480	0.0567	0.0567	0.0564	0.0442	0.0706	0.0705	0.0732	0.0714	0.0714	0.0705	**0.0509**	**0.0495**	**0.0682**	**0.0872**	**0.0681**	**0.0639**	**0.0323**

TABLE. V. THE PERFORMANCES OF ALTERNATIVES WITH RESPECT TO THE CRITERIA

	C11	**C12**	**C13**	**C21**	**C22**	**C23**	**C31**	**C32**	**C33**	**C51**	**C52**	**C53**	**C61**	**C62**	**C63**	**C71**	**C72**	**C73**
A1	0.9	0.3	1	1	2	1	3	1	2	2	3	0	2	180	2	1	2	{HIPAA, ISO 27001}
A2	2	0..5	1	1	3	0	2	0	3	3	2	1	3	230	3	1	2	{HIPAA, ISO 27001}
A3	1	0.6	1	1	3	1	4	1	3	1	4	1	3	356	3	1	3	{HIPAA, ISO 27001, SOC, PCI}
A4	0.9	0.1	1	1	3	1	4	0	3	1	4	1	2	365	4	1	4	{HIPAA, ISO 27001, SOC, PCI}
A5	1	0.5	1	1	2	1	3	1	2	2	3	0	2	230	2	1	3	{HIPAA, ISO 27001, SOC}

TABLE. VI. THE SUPERMATRIX

	Goal	C11	C12	C13	C21	C22	C23	C31	C32	C33	C51	C52	C53	C61	C62	C63	C71	C72	C73
Goal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C11	0.614	0	0.667	0.667	0.493	0.333	0.333	0.26	0.26	0.26	0.311	0.311	0.345	0	0	0	0	0	0
C12	0.117	0.667	0	0.333	0.196	0.333	0.333	0.328	0.328	0.328	0.196	0.196	0.109	0	0	0	0	0	0
C13	0.268	0.333	0.333	0	0.311	0.333	0.333	0.413	0.413	0.413	0.493	0.493	0.547	0	0	0	0	0	0
C21	0.345	0	0	0	0	0.75	0.667	0.493	0.493	0.493	0.249	0.249	0.359	0	0	0	0	0	0
C22	0.109	0	0	0	0.75	0	0.333	0.311	0.311	0.311	0.157	0.157	0.124	0	0	0	0	0	0
C23	0.547	0	0	0	0.25	0.25	0	0.196	0.196	0.196	0.594	0.594	0.517	0	0	0	0	0	0
C31	0.568	0.577	0.368	0.493	0.559	0.345	0.345	0	0.5	0.5	0.627	0.627	0.588	0	0	0	0.493	0.493	0.493
C32	0.075	0.081	0.082	0.311	0.089	0.109	0.109	0.8	0	0.5	0.094	0.094	0.089	0	0	0	0.311	0.311	0.311
C33	0.358	0.342	0.55	0.196	0.352	0.547	0.547	0.2	0.5	0	0.28	0.28	0.323	0	0	0	0.196	0.196	0.196
C51	0.311	0.493	0.333	0.493	0.4	0.196	0.493	0.594	0.594	0.594	0	0.5	0.5	0.493	0.493	0.493	0.493	0.493	0.493
C52	0.196	0.311	0.333	0.311	0.2	0.311	0.196	0.249	0.249	0.249	0.333	0	0.5	0.311	0.311	0.311	0.311	0.311	0.311
C53	0.493	0.196	0.333	0.196	0.4	0.493	0.311	0.157	0.157	0.157	0.667	0.5	0	0.196	0.196	0.196	0.196	0.196	0.196
C61	0.226	0	0	0	0	0	0	0	0	0	0.493	0.493	0.109	0	0.5	0.5	0	0	0
C62	0.101	0	0	0	0	0	0	0	0	0	0.311	0.311	0.163	0.5	0	0.5	0	0	0
C63	0.674	0	0	0	0	0	0	0	0	0	0.196	0.196	0.729	0.5	0.5	0	0	0	0
C71	0.14	0	0	0	0.14	0.196	0.169	0.328	0.328	0.328	0.126	0.126	0.493	0.379	0.493	0.493	0	0.5	0.25
C72	0.528	0	0	0	0.528	0.311	0.443	0.26	0.26	0.26	0.416	0.416	0.311	0.331	0.311	0.311	0.5	0	0.75
C73	0.333	0	0	0	0.333	0.493	0.387	0.413	0.413	0.413	0.458	0.458	0.196	0.289	0.196	0.196	0.5	0.5	0

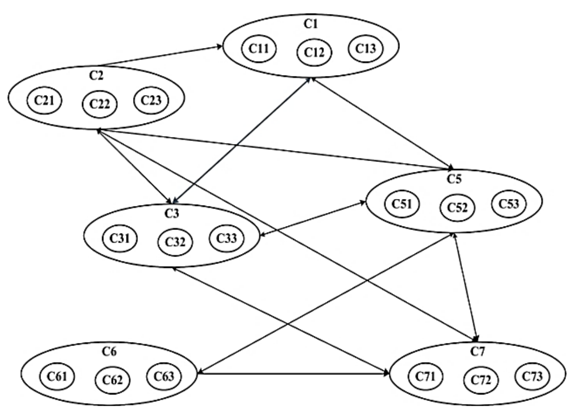


Fig. 4. Security Controls Network Structure based on Fuzzy DEMATEL Analysis.

Next, we calculate the weights of the control criteria to obtain the weighted supermatrix. As discussed in (step 2, section 3.2), the weight of a control criterion is the combination of its influence weight and the level of control granted to the user over its management. The influence weights of the control criteria are the α-cut total-influence sub-matrix for security controls. The level of control the user has over the criteria are defined as follows: $w_1^{cont} = 1, w_2^{cont} = 1, w_3^{cont} = 0.5, w_5^{cont} = 1,\ w_6^{cont} = 1, w_7^{cont} = 1$. The total weight of the cluster is the product of its influence degree and level of control. The clusters weights are used to calculate the weighted supermatrix and obtain the limit supermatrix to derive the total subjective weights of criteria. The subjective criteria weights are combined with objective weights following (steps 5-6, section 3.2). The coefficients are set to (α=β=0.5). The results are shown in Table VII.

TABLE. VII. THE OBJECTIVE, SUBJECTIVE, AND TOTAL CRITERIA WEIGHTS

Criteria		Subjective Weights	Objective Weights	Total Weights
Incident resolution	C11	0.0498	0.0326	0.0412
Incident response	C12	0.0418	0.0679	0.0548
Malware detection	C13	0.0691	0	0.0346
Data leakage prevention techniques	C21	0.0481	0	0.0241
Data deletion type	C22	0.0290	0.0103	0.0196
Encryption techniques	C23	0.0470	0.1255	0.0863
Authentication level	C31	0.1264	0.0161	0.0712
Third party authentication	C32	0.0478	0.2874	0.1676
Authentication mechanisms	C33	0.0748	0.0103	0.0425
VM encryption	C51	0.1320	0.0487	0.0903
Cryptographic hardware module protection level	C52	0.0766	0.0161	0.0463
Hypervisor access control policy	C53	0.0586	0.2874	0.1730
Log access availability	C61	0.0190	0.0115	0.0152
Logs retention period	C62	0.0129	0.021	0.0169
Network penetration tests	C63	0.0145	0.0199	0.0172
Independent audits	C71	0.0429	0	0.0215
Audit planning	C72	0.0525	0.0199	0.0362
Compliances	C73	0.0573	0.0255	0.0414

Next, we perform the ranking of alternatives following the TOPSIS method. The final results regarding the closeness distance to the ideal solution, and the final ranking are shown in Table VIII. The best alternative according to the results is (A3).

In the final phase, we perform the gap analysis for the best-selected alternative (A3) as discussed in section 3.4. The IGA map is shown in Fig. 5. We can further leverage the characteristics of DEMATEL to understand the cause-effect relationship between the different attributes based on the prominence level ($R_i + D_i$) and relation level ($R_i - D_i$) as discussed in (step 4, section 3.1). The relation level for the control criteria are as follows: ($R_i - D_i$: C1= -0.125, C2= 0.271, C3=0.119, C5=0.086, C6=0.153, C7=-0.083), calculated from the security control total influence matrix (Table IV). Both criteria C1 and C7 have a negative relation level, which means that they are effect criteria. The remaining criteria (C2, C3, C5, and C6) are cause criteria, representing the driving factors of the core problem. We plot the cause attributes into the IGA map following Eq. 14-15 (Section 3.4), as shown in Fig. 5. Most of the attributes fall into the "keep as it is" quadrant, while criterion C32 is considered a "neutral" attribute, and criteria C51 and C63 "critical" to improve. For example, the criterion C63 being a critical attribute, while most of the attributes being influenced fall into the "keep as it is". Then, the improvements towards this attribute should begin immediately along with the performance of the other attributes.

TABLE. VIII. THE DISTANCE MEASURES TO THE BEST IDEAL SOLUTION (S+), WORST SOLUTION (S-), CLOSENESS, AND THE FINAL RANKING OF ALTERNATIVES

	S+	S-	Closeness	Ranking
A1	0.1071	0.1091	0.5047	5
A2	0.1099	0.112	0.5047	4
A3	0.0445	0.151	0.7724	1
A4	0.1102	0.1129	0.5061	3
A5	0.1047	0.1112	0.5149	2

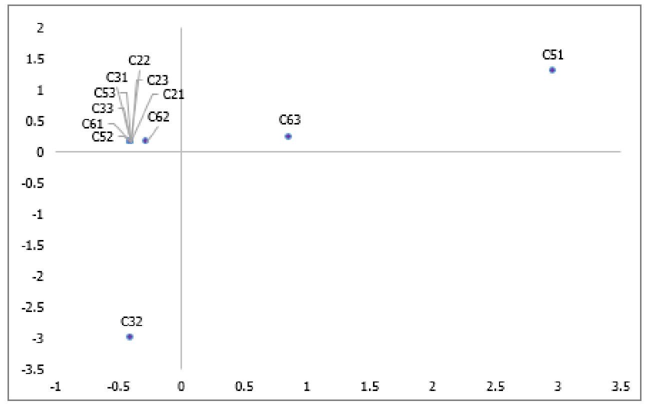


Fig. 5. The IGA Map based on the cause Attributes from DEMATEL for Best-Selected Alternative (A3).

V. CONCLUSION

In this paper, we proposed a holistic risk-driven security evaluation approach for cloud services selection. We addressed three main issues, namely, (1) lack of a systematic and quantitative approach for the selection of the minimal and representative criteria for cloud services security evaluation considering the dependency relations between cloud service models, the potential threats and vulnerabilities, and the effectiveness of the security controls; (2) lack of comprehensive criteria weighting approach considering the dependencies between control criteria and cloud stockholder's varying degree of control for implementing and managing the security services; and (3) lack of effectiveness-based evaluation for cloud services. The proposed method first builds the evaluation context and selects the core security controls (i.e., evaluation criteria) considering several factors, namely threat likelihood, vulnerability relevance, and controls effectiveness given the cloud service models using fuzzy DEMATEL method. Next, the weights of criteria were calculated based on the dependencies between the security controls, cloud user's level of control given the cloud service model and security control type, as well as user preferences using a combination of fuzzy DEMATEL and fuzzy ANP methods. Furthermore, subjective weights were combined with objective weights to obtain more reliable results. Finally, the TOPSIS method was employed for services ranking and the improvement gap analysis (IGA) method was leveraged to provide more insights on the strength and weaknesses of the selected services. The proposed method facilitates a systematic selection and prioritization of security controls for evaluation following a risk-driven approach, which drives for more efficient and effective services evaluation.

REFERENCES

[1] L. Badger, R. Patt-corner, and J. Voas, "NIST cloud computing synopsis and recommendations," Nist Spec. Publ., vol. 800, no. 146, p. 81, 2012.

[2] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing," J. Netw. Comput. Appl., vol. 34, no. 1, pp. 1–11, 2011.

[3] T. Halabi and M. Bellaiche, "A broker-based framework for standardization and management of Cloud Security-SLAs," Comput. Secur., vol. 75, pp. 59–71, 2018.

[4] H. Mohammad, A. Ahmad, and K. Bin, "A novel evaluation framework for improving trust level of Infrastructure as a Service," Cluster Comput., vol. 19, no. 1, pp. 389–410, 2016.

[5] J. Sidhu and S. Singh, "Improved TOPSIS method based trust evaluation framework for determining trustworthiness of cloud service providers," J. Grid Comput., vol. 15, no. 1, pp. 81–105, 2017.

[6] R. Krishankumar and K. S. Ravichandran, "Solving cloud vendor selection problem using intuitionistic fuzzy decision framework," Neural Comput. Appl., vol. 1, 2020.

[7] Z. Ma, R. Jiang, M. Yang, T. Li, and Q. Zhang, "Research on the measurement and evaluation of trusted cloud service," Soft Comput., vol. 22, no. 4, pp. 1247–1262, 2018.

[8] T. Halabi and M. Bellaiche, "Towards quantification and evaluation of security of Cloud Service Providers," J. Inf. Secur. Appl., vol. 33, pp. 55–65, 2017.

[9] The Cloud Service Measurement Initiative Consortium (CSMIC), "Service measurement index introducing the service measurement index (SMI)," pp. 1–8, 2011.

[10] R. Ranjan, K. Siba, and M. Chiranjeev, "A novel framework for cloud service evaluation and selection using hybrid MCDM methods," Arab. J. Sci. Eng., 2017.

[11] S. K. Garg, S. Versteeg, R. Buyya, "A framework for ranking of cloud computing services," Futur. Gener. Comput. Syst., vol. 29, no. 4, pp. 1012–1023, 2013.

[12] CSA. "Cloud Controls Matrix.". https://cloudsecurityalliance.org/group/cloud-controls-matrix.

[13] J. Luna, A. Taha, R. Trapero, and N. Suri, "Quantitative reasoning about cloud security using service level agreements," IEEE Trans. Cloud Comput., vol. 5, no. 3, pp. 457–471, 2017.

[14] J. Modic, R. Trapero, A. Taha, J. Luna, M. Stopar, and N. Suri, "Novel efficient techniques for real-time cloud security assessment," Comput. Secur., vol. 62, pp. 1–18, 2016.

[15] T.L. Saaty, "The analytic hierarchy process," McGraw-Hill, New York, (1980).

[16] T.L. Saaty, "The analytic network process," Pittsburgh: RWS, (1996).

[17] NIST, "NIST Cloud Computing security reference architecture," NIST Tech Beat, pp. 1–2, 2011.

[18] R. M. Savola and H. Pentikäinen, "Towards security effectiveness measurement utilizing risk-based security assurance," 2010 Inf. Secur. South Africa, pp. 1–8, 2010.

[19] G. Tontini and J. D. Picolo, "Improvement gap analysis," Manag. Serv. Qual., vol. 20, no. 6, pp. 565–584, 2010.

[20] S. Maroc and J. B. Zhang, "Risk-based and dependency-aware criteria specification for cloud services security evaluation," IEEE Inter. Conf. Comm. Soft and Netw. (ICCSN), 2019, pp. 731–735.

[21] E. Fontela, and A. Gabus, "DEMATEL innovative methods tech. report no. 2, structural analysis of the world problematique," NY: Battelle Geneva Resear. Inst., (1974).

[22] C. A. B. de Carvalho, R. M. de C. Andrade, M. F. de Castro, E. F. Coutinho, and N. Agoulmine, "State of the art and challenges of security SLA for cloud computing," Comput. Electr. Eng., vol. 59, pp. 141–152, 2017.

[23] H. Alabool, A. Kamil, N. Arshad, and D. Alarabiat, "Cloud service evaluation method-based Multi-Criteria Decision-Making: A systematic literature review," J. Syst. Softw., vol. 139, pp. 161–188, 2018.

[24] H. Mouratidis, S. Islam, C. Kalloniatis, and S. Gritzalis, "A framework to support selection of cloud providers based on security and privacy requirements," J. Syst. Softw., vol. 86, no. 9, pp. 2276–2293, 2013.

[25] H. M. Alabool et al., "A novel evaluation model for improving trust level of infrastructure as a service," 2015 Int. Symp. Math. Sci. Comput. Res. iSMSC 2015 - Proc., vol. 19, no. 1, pp. 162–167, 2016.

[26] N. Al-safwani, Y. Fazea, and H. Ibrahim, "ISCP : In-depth model for selecting critical," Comput. Secur., vol. 77, pp. 565–577, 2018.

[27] L. Sun, H. Dong, O. K. Hussain, F. K. Hussain, and A. X. Liu, "A framework of cloud service selection with criteria interactions," Futur. Gener. Comput. Syst., vol. 94, pp. 749–764, 2019.

[28] A. Taha, P. Metzler, R. Trapero, J. Luna, and N. Suri, "Identifying and utilizing dependencies across cloud security services," ASIA CCS 2016 - Proc. 11th ACM Asia Conf. Comput. Commun. Secur., pp. 329–340, 2016.

[29] C. Choi and H. Jeong, "Quality evaluation and best service choice for cloud," pp. 245–270, 2014.

[30] T. Subramanian and N. Savarimuthu, "Cloud service evaluation and selection using fuzzy hybrid mcdm approach in marketplace," vol. 5, no. 2, 2016.

[31] C. H. Su, G. H. Tzeng, and H. L. Tseng, "Improving cloud computing service in fuzzy environment - Combining fuzzy DANP and fuzzy VIKOR with a new hybrid FMCDM model," 2012 Int. Conf. Fuzzy Theory Its Appl. iFUZZY 2012, pp. 30–35, 2012.

[32] V. Ghafori and R. Manouchehri Sarhadi, "Best cloud provider selection using integrated ANP-DEMATEL and prioritizing SMI attributes," Int. J. Comput. Appl., vol. 71, no. 16, pp. 18–25, 2013.

[33] B. Alexandru, D. Mihaela, and D. Mihail, "A decision making framework for weighting and ranking criteria for Cloud provider selection," 2016.

[34] Y. O. Yang, H. Shieh, and G. Tzeng, "A VIKOR technique based on DEMATEL and ANP for information security risk control assessment," Inf. Sci. (Ny)., vol. 232, pp. 482–500, 2013.

[35] C. C. Lo and W. J. Chen, "A hybrid information security risk assessment procedure considering interdependences between controls," Expert Syst. Appl., vol. 39, no. 1, pp. 247–257, 2012.

[36] L.A. Zadeh, "Fuzzy set theory," Inf. and Contr. vol. 8, pp. 338–353, 1965.

[37] L. E. Quezada, H. A. López-ospina, P. I. Palominos, and A. M. Oddershede, "Identifying causal relationships in strategy maps using ANP and DEMATEL," Comput. Ind. Eng., vol. 118, no. January 2017, pp. 170–179, 2018.

[38] X.H. Xu, L.Y. Zhang, Q.F Wan, "A variation coefficient similarity measure and its application in emergency group decision-making," Syst. Eng. Proc. 5 119–124 (2012).

[39] K. Yoon, and C. L Hwang, "TOPSIS (Technique for order preference by similarity to ideal solution)-A multiple attribute decision making," (1980).

[40] J. A. Martilla and J. C. James, "Importance-Performance Analysis," J. Mark., vol. 41, no. 1, pp. 77–79, 1977.

Hybrid Machine Learning Algorithms for Predicting Academic Performance

Phauk Sokkhey[1]
Graduate School of Engineering and Science
University of the Ryukyus
1 Senbaru, Nishihara, Okinawa, 903-0123, Japan

Takeo Okazaki[2]
Department of Computer Science and Intelligent Systems
University of the Ryukyus
1 Senbaru, Nishihara, Okinawa, 903-0123, Japan

***Abstract*—The large volume of data and its complexity in educational institutions require the sakes from informative technologies. In order to facilitate this task, many researchers have focused on using machine learning to extract knowledge from the education database to support students and instructors in getting better performance. In prediction models, the challenging task is to choose the effective techniques which could produce satisfying predictive accuracy. Hence, in this work, we introduced a hybrid approach of principal component analysis (PCA) as conjunction with four machines learning (ML) algorithms: random forest (RF), C5.0 of decision tree (DT), and naïve Bayes (NB) of Bayes network and support vector machine (SVM), to improve the performances of classification by solving the misclassification problem. Three datasets were used to confirm the robustness of the proposed models. Through the given datasets, we evaluated the classification accuracy and root mean square error (RSME) as evaluation metrics of the proposed models. In this classification problem, 10-fold cross-validation was proposed to evaluate the predictive performance. The proposed hybrid models produced very prediction results which shown itself as the optimal prediction and classification algorithms.**

Keywords*—Student performance; machine learning algorithms; k-fold cross-validation; principal component analysis

I. Introduction

The poor performance of students in high school has become a worried-task for educators as it affects the secondary national exam and step to higher education. Mathematics is considered as the basic background for many science subjects, and give very strongly affect the national exam and for further study in higher education [1]. For example, students who are poor in mathematics are much more likely to fail in diploma national exams in Cambodia [2]. They later found themselves harder to choose a major for higher study and hard to survive in the university journey. Early prediction and classification of student performance level offers an early warning and gives a recipe for improving the poor performance of students as well as for other managerial settings. Hence, we aim to deal with the unknown behavior pattern of students which affects student performance. There are various factors affect the performance of students in mathematics; those factors consist of schooling factors, domestics or home factors, and personal or individual factors. These related factors were used as predictive features in predicting the achievement of students in mathematics.

In the age of the information revolution, analysis of the database in education environments such as learning analytics, predictive analytics, educational data mining, and machine learning techniques has become a hot area of research [3-5]. The supervised learning was used to predict, classify the students' performance and analyze their learning behaviors to follow up on their progress in classes. However, the challenging task is to find the optimal algorithm which could produce satisfying results. Machine learning algorithms such as naïve Bayes, logistic regression, artificial neural networks, decision tree, random forest, support vector machine, k-nearest neighbor, and more, were popularly used to analyze and predict academic performance [3-14]. The performance of each model is varied from dataset to dataset, which relies on the characteristics and quality of data.

In the classification problem, a reason for misclassification that declines the performance of the model is from the quality of data that disturbs the algorithms. Various literature has focused on using dimensional reduction (feature selection and feature extraction methods) to improve the prediction and classification performance. In our work, we applied principal component analysis (PCA) as a feature extraction technique to transform the original dataset into a new dataset of high quality. We also introduced 10-fold cross-validation is to evaluate the predictive performance of the models and to judge how they perform in a new dataset, the testing samples or test data.

This paper aims at proposing a novel hybrid approach of machine learning for solving the classification problem. The proposed hybrid approach is the combination of four baseline machine learning algorithms with 10-fold cross-validation and principal component analysis.

II. Related Works

Supervised learning in machine learning requires an effective prediction model for solving prediction and classification problems. As mentioned in the Introduction, the educational data mining (EDM) field has studied different machine learning techniques to determine these techniques obtaining a high accuracy to predict the future performance of students [3-5].

Table I summarized the popular and state-of-the-art classification algorithms, which were used to predict student performance in educational datasets. Several works have been investigated to find the best algorithms to predict future performance.

TABLE. I. SUMMARY OF COMMON MACHINE LEARNING CLASSIFIERS WHICH ARE USED IN PREDICTING STUDENT PERFORMANCE

Ref.	Main Results
[6]	(i) C4.5 and Randomtree were proposed. (ii) C4.5 could produce the highest accuracy.
[7]	(i) The six classifiers are decision tree (DT), random forest (RF), artificial neural network (ANN), Navie Bayes (NB), logistic regression (LR), and generalized linear model (GLM). (ii) The RF was found to be the best classifier.
[8]	(i) C4.5, NB, 3-nearest neighbor (3-NN), backpropagation (BP), sequential minimal optimization (SMO), LR were proposed, (ii) NB algorithms produced the highest classification result.
[9]	(i) Three tree-based classifiers: J48, Random Tree, and REPTree were used. (ii) J48 was found to be the best prediction model.
[10]	(i) NB, support vector machine (SVM), C4.5, CART are used to build the learning model. (ii) SVM is the best model compared to NB, C4.5, and CART.
[11]	(i) RF, multilayer perceptron (MLP), and ANN were used to classify student performance. (ii) The RF algorithms generated the highest accuracy.
[12]	(i) J48, CART, and RF classifiers were proposed with principal component analysis (PCA). (ii) PCA-RF was found to generate the highest accuracy.
[13]	(i) MLP, Radial Bias Function (RBF), SMO, J48, and NB are proposed to combine with PCA. (ii) PCA-NB generated the highest accuracy.
[14]	(i) Three Boosting algorithms (C5.0, AddaBoost M1., and AdaBoost SAMME) are proposed. (ii) The C5.0 outperformed the other two boosting models.

III. MACHINE LEARNING ALGORITHMS

We proposed hybrid models by a conjunction of machine learning algorithms with principal component analysis. We first proposed the baseline models. We then improved the performance of our proposed baseline models with k-fold cross-validation. Lastly, we proposed the hybrid machine learning model by combining it with principal component analysis as in Fig. 1.

A. The Baseline Models

There are numerous effective machine learning approaches that have been extensively applied to educational environments. For various purposes in educational settings, we need to take different machine learning techniques such as association rule mining, regression analysis, classification, and clustering [3]. Classification is a common technique in machine learning that was used in order to classify and predict the categories or predefined classes of target variables. In this work, we observed several machine learning classifiers and selected the four state-of-the-art methods which are popularly used in predicting academic performances [3-14]. The four proposed algorithms are support vector machine, naïve Bayes C5.0 of the decision tree, and random forest.

1) Support vector machine: A Support Vector Machine (SVM) is a kind of classification algorithm obtained by the mean of a separating hyperplane [15]. The concept of SVM is to create a line or a hyperplane to separates the samples into classes. SVM is used to observe for the optimal hypersurface to separate each two different data classes. Once the data is more complex, then we create more dimensional space to have a linear separation of data.

Given a training sample $(x_i, y_i),\ i = 1, 2, ..., m$, where $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$ are called the target classes, the classical SVM classifier is subject to solve the optimization problem:

$$\min_{w,b,\xi}\left\{\frac{1}{2}w^T w + C\left(\sum_{i=1}^{m}\xi_i\right)\right\}$$

$$\text{subject to: } y_i(w^T\phi(x_i)+b) \geq 1 - \xi_i,\ \xi_i > 0, \forall i, \tag{1}$$

where $\phi(x)$ is treated for nonlinear function case mapping x into a higher dimensional space. The parameters w, b and ξ_i represent the weight, bias, and slack variable, respectively. And the optimal hyperplane is possibly to be solved using Lagrangian and then transform it into a quadratic problem of the function $W(\alpha)$ as in (2):

$$\max\ W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i, x_j)$$

$$\text{subject to: } \sum_{i=1}^{m}\alpha_i y_i = 0; \alpha_i \in [0, \mathrm{C}], i = 1, 2, .., m, \tag{2}$$

where $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is the kernel function and, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_m)$ is a set of Lagrange multipliers.

The decision function can be written as:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{m}\alpha_i y_i K(x_i, x_j) + b\right). \tag{3}$$

Different kernel functions are used to help SVM to maximize margin hyperplanes to obtain the optimal solution. The most popular used kernels are the polynomial function, sigmoid function, and radial basis function. SVM with radial bias function (RBF) kernel is one of the most commonly used kernels for the multi-classification problem since it requires fewer parameters comparing to the polynomial kernel. Consequently, RFB is an appropriate choice to be used kernel. Hence, this work applied RBF as a kernel function top to get the optimal solution.

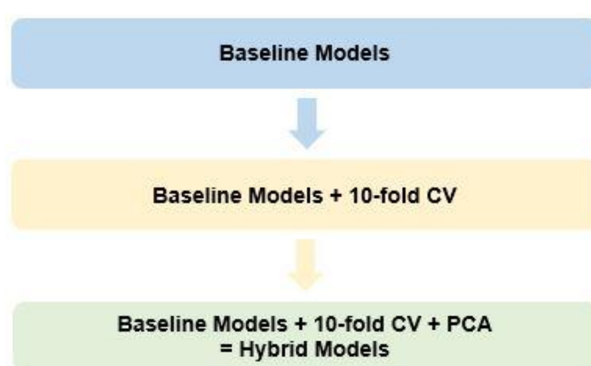


Fig. 1. Illustration of Task Procedure.

2) Naïve Bayes (NB): NB is one among the simple but effective machine learning algorithms that is preferably used in many classification problems. NB is a very attractive method for education research [16]. In the educational domain, an assumption of conditional independence is often ignored and disturbed. Considering that variables are inter-connected, the NB classifier can tolerate strong supervising dependence between independent variables. NB classifier is Bayes theorem-based method that used the idea of computing posterior probability for decision rule. NB classifier has been especially popular for educational data mining. Suppose D is a dataset of n dimensional vector X : $(x_1, x_2, x_3, ..., x_n)$ describing attributes of each student and suppose there are k classes: $C_1, C_2, ..., C_k$. NB classif-ier predicts X belong to a class C_i if and only if $P(C_i \mid X) > P(C_j \mid X)$ for all $1 \le j \le k, i \ne j$. The NB classifier is found on conditional Bayes probability as in (4):

$$P(C_i \mid X) = \frac{P(C_i) \times P(X \mid C_i)}{P(X)}. \tag{4}$$

The probability $P(X)$ is normalizing constant and $X = (x_1, x_2, ..., x_p)$ is the set of features variables with a strong assumption of independent predictors, then (4) can be rewritten as:

$$\begin{aligned} P(C_i \mid X) &\propto P(C_i)\mathrm{P}(\mathrm{X} \mid \mathrm{C}_i) \\ &= P(C_i)\prod_{j=1}^{n} P(x_j \mid C_i). \end{aligned} \tag{5}$$

The naïve Bayes classifier holds many advantages such as it is a very simple algorithm, not contain any parameter to optimize, efficient for classification, and easy to interpret.

3) C5.0: Decision tree is a "non-parametric white-box model" which is simple and effective for classification and regression tasks while C5.0 is one of the most famous algorithms of decision tree that construct the structure in the form of tree diagram [14]. This algorithm takes care of various of the decisions automatically using fairly reasonable defaults.

C5.0 is a successor of C4.5; it builds tree structure from training set using the idea of Shannon entropy. The algorithm purifies the subset of samples via the concept of information entropy. Entropy defines the impurity of any subset of an sample set S at a specific node N is written as:

$$Entropy(S) = I(S) = -\sum_{i=1}^{c} P(c_i)\log_2 P(c_i). \tag{6}$$

The constant c is denoting the number of classes and $P(c_i)$ is the proportion of values in the class i . After obtaining the measure of purity, the algorithm needs to decide which feature to split next. The algorithm calculates homogeneity resulting from a split on each possible feature, this procedure of calculation is called information gain (IG) as shown in (7):

$$IG(S, A) = I(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} I(S_v). \tag{7}$$

One complicated matter after splitting is that a split result in more than one partition that is what we need to compute what is called split information in the following equation:

$$SplitInfo_A(S) = -\sum_{i=1}^{v} \frac{|S_i|}{|S|} \log_2\left(\frac{|S_i|}{|S|}\right). \tag{8}$$

Then, using information gain as see formula (7), and splitting information as in (8), we then can compute the information gain ration using the following equation:

$$GainRatio(S, A) = \frac{IG(S, A)}{SplitInfo_A(S)}. \tag{9}$$

The C5.0 of a decision tree is one of the most popular machine learning algorithms that has been widely used in various applications.

4) Random Forest (RF): As in the name indicates its meaning, the random forest is an algorithm builds the forest with a number of trees. A random forest algorithm is a tree-based tool that grows many classification trees [12]. It is a kind of ensemble classifier that combines several classification trees to create a new classifier. The concepts of bootstrap aggregation or bagging method is used to grow each tree. To classify a new example, each decision tree gives a classification for the input data which is so-called “voting for a class”. The RF algorithm chooses a class with the highest votes. The illustration of the process of random forest algorithms is shown in Fig. 2.

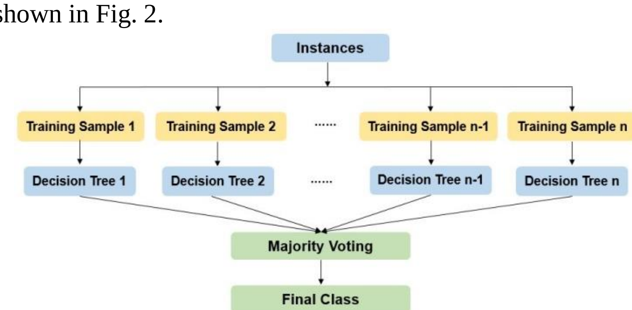


Fig. 2. Illustration of Random Forest Algorithm.

B. The k-fold Cross-Validation

Cross-validation is one of statistical technique that used to test the effectiveness of machine learning algorithms. There are various methods of cross-validation but the k-fold cross-validation is chosen since it is popular and easy to understand, also generally generates a lower bias comparing to the other cross-validation methods. The process of k-fold cross-validations is summarized as the following:

1) Shuffle the entire samples randomly
2) Split samples into k sub folds
3) In the split k sub folds:

- Take 1 fold as a holdout or test set

- Take the remaining $k-1$ folds as the training set
- Retain the evaluation score and discard the model

4) Repeat the iteration until every single fold was treated as a testing set. Finally, compute the average score of the recorded scores.

In our study, we chose the 10-fold cross-validation (will be shortly called 10-CV) to access our proposed algorithms. This process is precisely illustrated in Fig. 3.

C. The Proposed Hybrid Models

The majority task in supervised machine learning is classification. The classification problem is a hot issue in data mining and machine learning. We proposed the four most popular classifiers that hold many merits. However, the major problem for those classifiers is overfitting and noisy data which leads to misclassification and deduce the accuracy of the classification. To overcome this matter, we try to reduce irrelevant feature and non-correlated features which disturb in the classification process. In data analysis, it requires more computational resources and consumes much time when that data consists of a huge volume. Hence, the feature extraction approach to remove noises in data in order to reduce time and resource usage and regain the high quality of data. The dimensional reduction could improve accuracy and boost up the performance by combining it with classification techniques. Using more high-quality data and feature reduction is one of the effective approaches to improve the performance of machine learning models. The four proposed models: support vector machine using radial basis function kernel (SVMRBF), naïve Bayes (NB), decision tree C5.0, and random forest (RF) are the affective algorithms for the classification problem, yet there is no perfect algorithm in machine learning.

SVM is a classifier with the use of support vectors called hyperplanes to separate data into classes. Thus, for a high dimensional dataset, the input space is high and can be unclean which is mostly declining the performance of the SVM algorithm. Thus, it requires an effective feature extraction method that discards noisy, irrelevant and redundant data, and still contains the useful information of data. Removal of such features can increase the search speed and accuracy rate.

NB is a classifier that holds many advantages, yet the greatest weakness of the NB classifier is that it relies on the often-faulty assumption of equally important and independent features. If there are any features that are irrelevant to some class C_k then the whole probability goes to zeros for that class because of production in equation (5), which leads to misclassification. In order to solve this problem, feature extraction will be the best tool to reduce irrelevant features and also improve the classification performance.

In the tree-based algorithms C5.0 and RF, the major problem in the splitting process of the decision tree is overfitting. Overfitting caused by noisy data and irrelevant features that produce misclassification results. In return, overfitting lowering the accuracy of tree-based classifiers. To reduce high dimensional data which, contains noisy and irrelevant data, a commonly-used technique is to use feature extraction in order to obtain a lower-input space that contains relevant and informative input features.

In order to improve the performance of the proposed machine learning algorithms, we proposed commonly-used feature extraction approach: principal component analysis (PCA) in this study. PCA is a statistical method that transforms an original data set to a new dataset of a lower dimension. The original dataset consists of possibly correlated variables are converted into a set of linearly uncorrelated variables.

PCA is one of the most popular dimensionality reduction algorithm [17]. In the PCA procedure, the data is first transformed into standardized data with zero mean. The idea behind getting the principle components is the covariance matrix is computed in order to obtain eigenvector and eigenvalues. The eigenvector with the highest eigenvalue is treated as the principal component of new data which shows the most significant relationship of input feature. PCA is less sensitive to different datasets than other holistic methods, so it is the most widely used technique as one of the effective feature reduction methods.

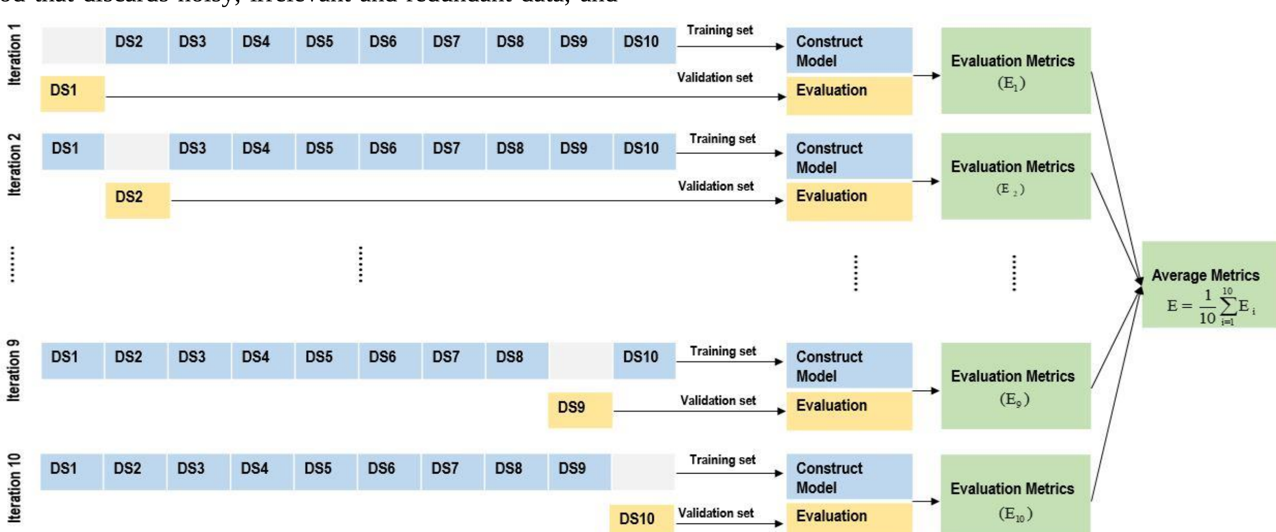


Fig. 3. Illustration of the K-Fold Cross-Validation Algorithm.

The procedure of transforming original dataset X of l dimension consisting of possibly correlated features to a new dataset Z of lower dimension $m(\mathrm{m} < l)$ consisting of linear uncorrelated features is as follows:

1) Compute mean: From the already processed data, first, find the mean of each attribute using the equation:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10}$$

2) Compute variance: In order to investigate and deviation of each feature in the dataset, we compute the variance using equation (11):

$$Var(X) = \sigma_x^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{11}$$

3) Compute covariance: Given two variables, denoted X and Y, the covariance and correlation are calculated using equation (12):

$$Cov(X,Y) = \sigma_x^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y) \tag{12}$$

$Cov(X,Y)$ equals to zero means that the two attributes X and Y are independent. Using equation (11) and (12), we can obtain covariance matrix S, which the entry s_{ij} , $i \neq j$, is the covariance between the i^{th} and j^{th} variables, and diagonal s_{ii} is the variance of i^{th} variables.

4) Compute Eigenvalues and Eigenvectors: The features in the new datasets are characterized by mean of eigenvectors and eigenvalues. The obtained eigenvectors will tell the direction of new features space while the eigenvalues are its magnitude. The eigenvalues are possible to obtain by solving the equation:

$$Det(S - \lambda I) = 0, \tag{13}$$

where the covariance matrix S is symmetric, λ is the eigenvalue of the symmetric matrix S , and I is an identity matrix. The eigenvector v corresponding to each eigenvalue λ can be computed via the equation:

$$(S - \lambda I)v = 0 \tag{14}$$

We denoted $E = \{v : (S - \lambda I)v = 0\}$ as the Eigen space containing all eigenvectors.

5) Obtain orthonormal eigenvectors: By means of linear algebra concept, we can obtain the nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq ...\lambda_m > 0$ with corresponding orthonormal eigenvectors $v_1, v_2, ..., v_m$. The eigenvectors are called the principal components of the dataset.

The proposed hybrid models by conjunction machine learning models with PCA are introduced for predicting and classifying the academic performance. The best benefits of PCA are summarized as follow:

a) Removing the high noises from samples and uncorrelated features from the collected dataset in the preprocessing step.

b) Reducing the high dimensional data to low dimensional one which remains the important characteristics of data that reduce overfitting problems.

c) Enhance the equality of features by getting rid of correlated features that effectively improve the performance of classification.

In this proposed research, we proposed the hybrid models by a conjunction of four baseline models (SVMRFB, NB, C5.0, and RF) with 10-fold cross-validation (10-CV) and principal component analysis (PCA).

IV. Datasets and Preprocessing

A. Datasets

In our study, we tried to collect all unseen features affecting student performance in mathematics subjects. Datasets contained 43 features describing the information of the learning behaviors of each student and one target variable describing the performance levels of students based on their score. The predictive features consist of the features observing from three main affected factors. These main factors contain the forty-three variables and their descriptions are shown in Table II. Table III described the predefined classes of the target variable.

To confirm the robustness and effectiveness of our proposed algorithms, we used three datasets. The first two datasets are generated datasets namely GDS1 (2000 samples) and GDS2 (4000 samples) that were constructed based on proposed structures of predictive features to the output variable as stated in [18-20]. The third dataset is the actual dataset that was collected from 22 high schools in Cambodia. The data collection was made using questionnaires form. Students were asked to provide their demographical information related to external effects such as domestic factors, individual or student factors, and school factors. The score of mathematics of students in the semester I was obtained from the administrative offices in each school. The dataset was named ADS3 that consists of 1204 samples.

TABLE. II. The Output Variable

N	Performance Levels	Score-based discretization
1	Excellent learner	90% and above
2	Good learner	75% to less than 90%
3	Average learner	60% to less than 75%
4	Slow learner	Less than 60%

TABLE. III. THE FACTORS AFFECTING STUDENT PERFORMANCE IN MATHEMATICS

N	Variables	Description	Type
Domestic Factors			
1	PEDU1	Father's educational level	Nominal
2	PEDU2	Mother's educational level	Nominal
3	POCC1	Father’s occupational status	Nominal
4	POCC2	Mother’s occupational status	Nominal
5	PSES	Family’s socioeconomic	Ordinal
6	PI1	Parents’ attention to students’ attitude	Ordinal
7	PI2	Parents’ time and money spending	Ordinal
8	PI3	Parents’ involvement as education	Ordinal
9	PS1	Parents’ feeling responsive and need	Ordinal
10	PS2	Parents’ respond to children’s attitude	Ordinal
11	PS3	Parents’ encouragement	Ordinal
12	PS4	Parents’ compliment	Ordinal
13	DE1	Domestic environment for study	Ordinal
14	DE2	Distance from home to school	Nominal
Student or Individual Factors			
15	SELD1	Number of hours for self-study	Nominal
16	SELD2	Number of hours for private math study	Ordinal
17	SELD3	Frequency of doing math homework	Ordinal
18	SELD4	Frequency of absence in math class	Ordinal
19	SELD5	Frequency of preparing for the math exam	Ordinal
20	SIM1	Student’ s interest in math	Ordinal
21	SIM2	Student’s enjoyment in math class	Ordinal
22	SIM3	Student’s attention in math class	Ordinal
23	SIM4	Student’s motivation to succeed in math	Ordinal
24	ANXI1	Student’s anxiety in math class	Ordinal
25	ANXI2	Student’s nervous in the math exam	Ordinal
26	ANXI3	Student’s feeling helpless in math	Ordinal
27	POSS1	Internet’s use at home	Binary
28	POSS2	Possession of computer	Binary
29	POSS3	Student’s study desk at home	Binary
School Factors			
30	CENV1	Classroom environment	Ordinal
31	CU1	Content’s language in math class	Nominal
32	CU2	Class session	Nominal
33	TMP1	Teacher mastering in math class	Ordinal
34	TMP2	Teacher’s absence in math class	Ordinal
35	TMP3	Teaching methods in math class	Ordinal
36	TMP4	Teacher’s involving in education’s content	Ordinal
37	TAC1	Math teacher’s ability	Ordinal
38	TAC2	Teacher’s encouragement to students	Ordinal
39	TAC3	Math teacher’s connection with students	Ordinal
40	TAC4	Math teacher’s help	Ordinal
41	ARES1	Adequate number of math teacher	Nominal
42	ARES2	Adequate use of classroom	Nominal
43	ARES3	Adequate use of math handout	Nominal

B. Preprocessing Tasks

Data preprocessing is an integral step in data mining that is used to transform the raw dataset into a clean and executable format to be ready for implementation. The preprocessing step is not only used to ensure the readiness of data suitable and ready for modeling but also to improve the performance of the models. The preprocessing tasks in this study contain some operations such as data cleaning or cleansing, data transformation, and data discretization. During data collection, the questionnaire completion was done with missing some questions and inputting invalid value (outliers). In our datasets, the number of missing values is low, so we used the imputing method in order to clean our data. We replaced the missing value in our categorical variables by its modes or high frequency-category values. In the output variable, there is a few missing value and outliers, then we replaced it by the mean value. For simplicity, we transformed some numerical features into ordinal types. In our study, we also discretized the output variables into four performance levels as shown in Table I.

V. EVALUATION METRICS

The performance of each proposed model in analyzing and predicting student performance can be evaluated from the analysis of the graphical confusion matrix. Without loss of generality, our output variable can be categorized into four ordinal categories as mention in Table I. Table IV shows the graphical confusion matrix which represents four classes of student performance level in mathematics subject. Class 1 presents the highest class, Class 2 denotes the second upper class, Class 3 describes the third class lower, and Class 4 denotes the lowest (poor) group of students. The below parameters are calculated.

A. Classification Accuracy

Accuracy is used to quantify the percentage of correctly predicted. Here, we want to evaluate the potential of our prediction model by measuring the percentage of correctly predicted the level of student performance as in (15):

$$Accuracy = \frac{\sum a_{ii}}{\sum a_{ij}} \times 100\% \tag{15}$$

B. Root Mean Square Error (RMSE)

We aim not only to predict the ability of students' performance levels but also to estimate how much our prediction is close to their performance level. We encoded these ordinal performance levels {slow, average, good, excellent} as {1,2,3,4}, respectively. The RMSE can be computed as:

$$RMSE = \sqrt{\sum_{i=1}^{M} \frac{(Pl_i^a - Pl_i^p)^2}{M}} \tag{16}$$

where $Pl^a \in \{1,2,3,4\}$ is the actual performance level and $Pl^p \in \{1,2,3,4\}$ is the predicted performance level. Contrasting with accuracy, the smaller the RMSE, the better the model is. RMSE equal to 0 shows the prediction model is perfect.

TABLE. IV. GRAPHICAL CONFUSION METRIC

		Predicted Classes			
		Slow	Average	Good	Excellent
Actual Classes	Slow	a_{11}	a_{12}	a_{13}	a_{14}
	Average	a_{21}	a_{22}	a_{23}	a_{24}
	Good	a_{31}	a_{32}	a_{33}	a_{34}
	Excellent	a_{41}	a_{42}	a_{43}	a_{44}

VI. EXPERIMENTAL RESULTS

In our experiments, we proceed in three phases. Phase 1 is to implement for the result of the baseline models. Phase 2 is to improve the baseline models by 10-fold cross-validation (10-CV). Phase 3 is to execute a hybrid model which is the combination of the baseline models with 10-CV and PCA.

A. Result of Baseline Models

We proposed four most popular machine learning techniques, random forest (RF), C5.0 of the decision tree, support vector machine using radial basis function kernel (SVMRBF), and naïve Bayes (NB) of the Bayesian network. The two performance metrics, classification accuracy, and RMSE are shown in the tables.

From Table V, VI, and VII, NB was found to be the poorest model, while the RF technique generates the highest performance with respect to both classification accuracy and RMSE, which shown itself as the potential model.

B. Results of Baseline Models with k-fold Cross-Validation

The k-fold cross-validation is a technique that is popularly used in prediction and classification models to split the dataset into $k-1$ sub folds for training and 1 fold for testing sets, then rotate the folds. In this experiment, we used 10-fold cross-validation, since it performs best at this split. 90% of the data was used in the training section, and 10% was used for testing purposes as shown in Fig. 3. Lastly, when all interactions were done, an average of all evaluation metrics is computed.

From Table VIII, the accuracy of SVMRBF was improved by 2%. The performance of the poor NB classifier was then much improved by to 68.03%. The 10-CV technique improved C5.0 and RF with an accuracy increase of 27% and 15%, respectively.

From Table IX, by shuffling the dataset GDS2 with 10-CV, the accuracy of SVMRBF algorithm was improved from 75.52% to 91.15%, which is a very good improvement. NB increased by an accuracy of 9%. The tree-based classifiers C5.0 and RF were improved by the accuracy of 9% and 6%, respectively.

TABLE. V. PERFORMANCE OF BASELINE MODELS TO GDS1

Bassline Models	Accuracy	RMSE
SVMRBF	75.01%	0.516
NB	35.79%	1.191
C5.0	78.42%	0.487
RF	80.06%	0.431

TABLE. VI. PERFORMANCE OF BASELINE MODELS TO GDS2

Bassline Models	Accuracy	RMSE
SVMRBF	75.52%	0.489
NB	67.68%	0.664
C5.0	86.18%	0.372
RF	90.37%	0.321

TABLE. VII. PERFORMANCE OF BASELINE MODELS TO ADS3

Bassline Models	Accuracy	RMSE
SVMRBF	86.44%	0.823
NB	65.02%	1.016
C5.0	76.55%	0.845
RF	89.23%	0.516

TABLE. VIII. PERFORMANCE OF BASELINE MODELS AND BASELINE MODELS+10-CV TO GDS1

Models	Accuracy	RMSE
SVMRBF	75.01%	0.516
SVMRBF + 10-CV	77.08%	0.456
NB	35.79%	1.191
NB+ 10-CV	68.03%	0.654
C5.0	78.42%	0.487
C5.0+ 10-CV	95.24%	0.185
RF	80.06%	0.431
RF+ 10-CV	96.48%	0.143

TABLE. IX. PERFORMANCE OF BASELINE MODELS AND BASELINE MODELS+10-CV TO GDS2

Models	Accuracy	RMSE
SVMRBF	75.52%	0.489
SVMRBF + 10-CV	94.15%	0.274
NB	67.68%	0.664
NB+ 10-CV	76.47%	0.498
C5.0	86.18%	0.372
C5.0+ 10-CV	95.69%	0.174
RF	90.37%	0.321
RF+ 10-CV	96.58%	0.139

TABLE. X. PERFORMANCE OF BASELINE MODELS AND BASELINE MODELS+10-CV TO ADS3

Models	Accuracy	RMSE
SVMRBF	86.44%	0.823
SVMRBF + 10-CV	90.66%	0.678
NB	65.02%	1.016
NB+ 10-CV	92.44%	0.145
C5.0	76.55%	0.845
C5.0+ 10-CV	94.82%	0.114
RF	89.23%	0.561
RF+ 10-CV	98.22%	0.113

From Table X, the NB accuracies improved rapidly from 65.44% to 90.66%. SVMRBF could yields around 4% better than the previous baseline SVMRBF. C5.0 and RF are tree-based classifiers that could produce a high risk of over-fitting. With a 10-CV, we can not only obtain better performance but also avoid overfitting problems too. By mean of 10-CV, accuracies of C5.0 and RF were improved to 94.82% and 98.22% which improved 18% and 9%, respectively.

C. Results of Proposed Hybrid Models

Our proposed hybrid models were constructed by combing the baseline models with a feature reduction approach, PCA. Feature extraction is one of the powerful methods in classification models that are used for the purpose of removing irrelevant or non-related features. Dimensionality reduction via PCA [13] can definitely serve as regularization in order to prevent overfitting and improve the model accuracies. Often, people end up making a mistake in thinking that PCA selects some features out of the dataset and discards others. The algorithm actually constructs a new dataset of properties based on a combination of the old ones.

In this section, we proposed the hybrid models as the combination of 10-CV in the previous section to PCA in order to avoid overfitting and more improvement in predicting performance. Tables XI, XII, and XIII describe the results of the proposed models to the three datasets, GDS1, GDS2, and ADS3, respectively.

We visualized the performance of the proposed models to the three datasets GDS1, GDS2 and ADS3 in Fig. 4, 5, and 6, respectively. In Fig. 4, the accuracy based in dataset GDS1, our proposed hybrid models boost the accuracy of SVMRBF from 75.01% to 83.88%, NB from 35.79% to 86.27%, C5.0 from 78.42% to 98.32%, and RF from 80.06% to 98.92%.

In Fig. 5, the hybrid models improved SVMRBF, NB, C5.0, and RF with accuracies of 20%, 23%, 12%, and 9%, respectively. In Fig. 6, the proposed hybrid SVMRBF could improve the classification accuracy from 86.44% to 97.01%. Classification through NB could yields 30% better than baseline NB. The accuracies of C5.0 and RF were improved to 99.25% and 99.72% correctly classified.

TABLE. XI. PERFORMANCE OF BASELINE MODELS, BASELINE MODELS +10-CV, AND HYBRID MODELS TO GDS1

Models	Average Accuracy	Lowest Accuracy	Highest Accuracy	Std.	Average RMSE	Lowest RMSE	Highest RMSE	Std.
SVMRBF	75.01%	70.27%	77.01%	1.421	0.516	0.460	0.691	0.059
SVMRBF+10-CV	7.08%	75.67%	78.89%	1.124	0.496	0.456	0.524	0.024
Hybrid SVMRBF	**83.88%**	**82.01%**	**85.05%**	**1.123**	**0.414**	**0.396**	**0.437**	**0.016**
NB	35.79%	32.41%	37.27%	1.861	1.191	1.045	1.411	0.127
NB+10-CV	68.03%	66.61%	69.82%	1.363	0.645	0.577	0.768	0.070
Hybrid NB	**86.27%**	**83.40%**	**90.35%**	**2.695**	**0.521**	**0.456**	**0.608**	**0.060**
C5.0	78.42%	75.41%	82.72%	2.429	0.487	0.449	0.543	0.038
C5.0+10-CV	95.24%	93.18%	96.28%	0.806	0.185	0.158	0.242	0.026
Hybrid C5.0	**98.32%**	**97.18%**	**99.28%**	**0.564**	**0.067**	**0.043**	**0.145**	**0.027**
RF	80.06%	77.25%	83.21%	1.860	0.431	0.371	0.495	0.037
RF+10-CV	96.48%	95.21%	97.52%	0.764	0.143	0.122	0.189	0.015
Hybrid RF	**98.92%**	**97.06%**	**99.78%**	**0.817**	**0.056**	**0.031**	**0.126**	**0.026**

TABLE. XII. PERFORMANCE OF BASELINE MODELS, BASELINE MODELS+10-CV, AND HYBRID MODELS TO GDS2

Models	Average Accuracy	Lowest Accuracy	Highest Accuracy	Std.	Average RMSE	Lowest RMSE	Highest RMSE	Std.
SVMRBF	75.52%	70.00%	77.80%	1.606	0.489	0.449	0.524	0.023
SVMRBF+10-CV	94.15%	92.89%	95.51%	0.909	0.274	0.234	0.347	0.050
Hybrid SVMRBF	**96.32%**	**95.52%**	**96.89%**	**0.591**	**0.182**	**0.173**	**0.202**	**0.011**
NB	67.68%	65.01%	69.82%	1.614	0.664	0.584	0.768	0.059
NB+10-CV	76.47%	73.52%	78.21%	1.624	0.498	0.466	0.550	0.031
Hybrid NB	**91.42%**	**88.71%**	**95.05%**	**1.593**	**0.321**	**0.288**	**0.383**	**0.033**
C5.0	86.18%	84.45%	88.41%	1.454	0.372	0.319	0.533	0.059
C5.0+10-CV	95.69%	93.28%	97.28%	1.026	0.174	0.141	0.197	0.017
Hybrid C5.0	**98.62%**	**98.18%**	**99.54%**	**0.475**	**0.067**	**0.043**	**0.145**	**0.028**
RF	90.37%	89.01%	91.80%	1.021	0.321	0.286	0.345	0.018
RF+10-CV	96.58%	95.21%	98.50%	0.928	0.139	0.114	0.189	0.016
Hybrid RF	**99.08%**	**97.60%**	**99.80%**	**0.732**	**0.057**	**0.031**	**0.126**	**0.027**

TABLE. XIII. PERFORMANCE OF BASELINE MODELS, BASELINE MODELS+10-CV, AND HYBRID MODELS TO ADS3

Models	Average Accuracy	Lowest Accuracy	Highest Accuracy	Std.	Average RMSE	Lowest RMSE	Highest RMSE	Std.
SVMRBF	86.44%	81.06%	90.69%	2.56	0.823	0.691	1.016	0.089
SVMRBF+10-CV	90.66%	86.66%	95.00%	2.86	0.678	0.364	0.813	0.131
Hybrid SVMRBF	**97.01%**	**95.34%**	**98.67%**	**1.112**	**0.178**	**0.114**	**0.230**	**0.042**
NB	65.02%	60.24%	69.21%	2.961	1.016	0.834	1.331	0.164
NB+10-CV	94.82%	91.18%	96.27%	1.165	0.154	0.133	0.232	0.035
Hybrid NB	**98.94%**	**98.01%**	**99.69%**	**0.607**	**0.145**	**0.042**	**0.230**	**0.049**
C5.0	76.55%	70.43%	80.73%	2.851	0.845	0.703	0.991	0.097
C5.0+10-CV	97.54%	94.79%	99.50%	1.923	0.114	0.070	0.160	0.034
Hybrid C5.0	**99.25%**	**98.21%**	**100%**	**0.606**	**0.073**	**0.000**	**0.145**	**0.045**
RF	89.23%	86.71%	92.35%	1.566	0.561	0.411	0.667	0.066
RF+10-CV	98.22%	95.69%	99.52%	1.353	0.113	0.070	0.160	0.034
Hybrid RF	**99.72%**	**99.01%**	**100%**	**0.357**	**0.041**	**0.000**	**0.077**	**0.029**

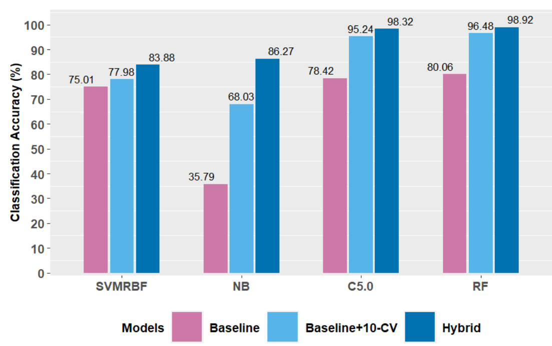


Fig. 4. Performance-based on the accuracy of GDS1

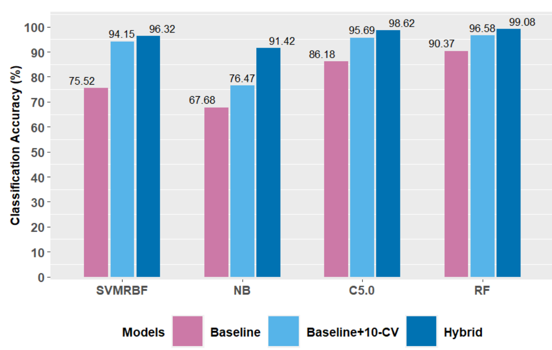


Fig. 5. Performance-based on the accuracy of GDS2

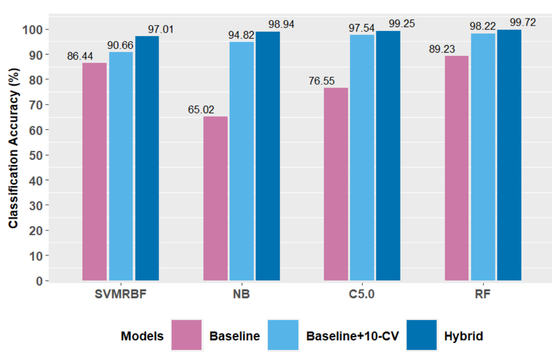


Fig. 6. Performance-based on the accuracy of ADS3

Fig. 7, 8, and 9 demonstrated the performance based on the accuracy of each model via each phase. We found the improvement by using 10-CV combined with PCA gives the best result in predicting student performance. The figures show the performance of the RMSE of the models in each step. The proposed hybrid models could generate a very small RMSE. The hybrid RF algorithm produced the smallest value of RMSE which shows itself as the best predictive model in this prediction problem.

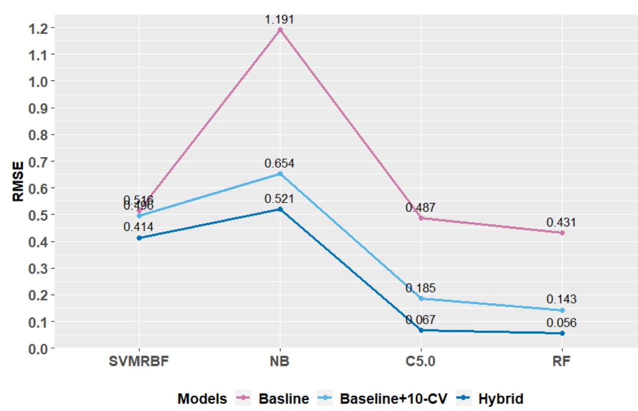


Fig. 7. Performance-based on the RMSE of GDS1.

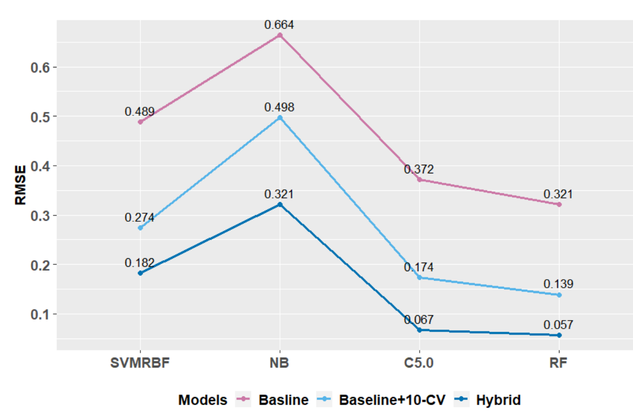


Fig. 8. Performance-based on the RMSE of GDS2.

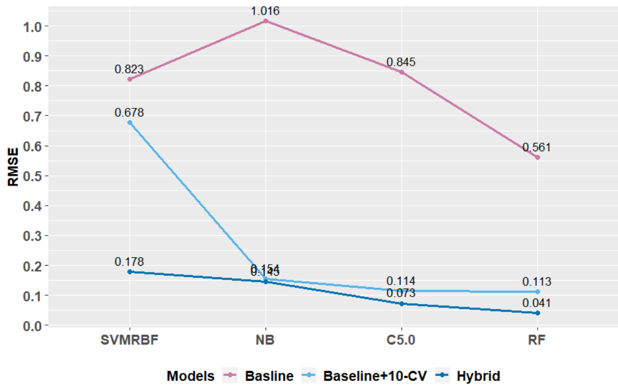


Fig. 9. Performance-based on the RMSE of ADS3.

From the results, by using 10-CV, we can improve the performance of our baseline models. Additionally, we observed that the proposed novel hybrid models could boost up the classification performance to the superior results. This proposed hybrid models can be regarded as an optimal prediction models for solving prediction and classification problems.

VII. Conclusion

This paper introduced the four popular classifiers of machine learnings to predict student performance. The four proposed algorithms are SVMRBF, NB, C5.0, RF. The procedure was made with three phases. Firstly, we observed the performance of those baseline methods. Secondly, we improved the performance with 10-CV. Lastly, we combined the PCA to baseline models, and 10-CV method to improve the classification performance. Based on classification accuracy and RMSE as measurement parameters, it shows that the proposed hybrid models by conjunction of the proposed models with PCA and 10-CV produced very satisfying results. In conclusion, by combining the baseline models with principal component analysis, and evaluated by k-fold cross-validation, the proposed hybrid models produced a high performance which shows itself as a potential algorithm for solving prediction and classification problem.

References

[1] Herbert K., "The New Book of Popular Science", World Applied Sciences Journal, Daribury, Connecticut: Grolier Inc., 1978.

[2] Ministry of Education, Youth and Sport, “Education in Cambodia: Finding from Cambodia's Experience in PISA for Development”, Phnom Penh: Author, 2018.

[3] S. Slater, S. Joksimovic, V. Kovanovic, R.s Baker, and D. Gasevic, "Tools for Educational Data Mining: A Review", Journal of Educational and Behavioral Statistics, Vol. 42, No. 1, 2016, pp. 88-106.

[4] Pooja Thakar, Anil Mehta, and Manisha, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue", International Journal of Computer Application, Vol. 100, No.12, January 2015, pp. 60-68.

[5] C. Romero and Ventura., “Educational Data Mining: A Review of the State of Art”, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 40, No. 6, 2010, pp. 601-618.

[6] Akinrotimi A.O, and Aremu D.R, "Student Performance Prediction Using Randomtree and C4.5," Journal of Digital Innovation and Contemporary Research, Engineering and Technology, Vol. 6, No. 3, 2018, pp. 23-34.

[7] Amjad A. S., Mostafa Al-Emran, and Khaled S., “Mining Student Information System Records to Predict Students’ Academic Performance”, Springer Nature Switzerland AG 2020, AMLTA 2019, AISC 921, 2019, pp. 229-239.

[8] Kotsiantis S., Piarrekeas C., and Pintelas P., "Predicting Students' performance in Distance Learning using Machine Learning Techniques", Applied Artificial Intelligence, Vol. 18, 2007, pp. 411-426.

[9] Hamoud A. K., Hashim A. S., and Awadh W. A., "Predicting Student Performance in Higher Education Institutes Using Decision Tree Analysis", International Journal of Interactive Multimedia and Artificial Intelligent, Vol. 5, No. 2, February 2018, pp. 26-31.

[10] Daud A., Aljonhani N.R., Abbasi R.A., Lytras M.D., Abbas F., Alowibdi J.S., “Prediction student performance using advanced learning analytics”, Proceedings of 26th International Conference on World Wide Web, Companion, Perth, Australia, April 2016, pp. 416-421.

[11] M.S. Mythili,, A.R.M. Shanavas, “An Analysis of Students' Performance using Classification Algorithms”, IOSR Journal of computer Engineering (IOSR-JCE), Vol 16, No. 1, January 2014, pp. 63-69.

[12] Aung Nway Oo, “Comparative Study of Principle Component Analysis based on Decision Tree Algorithm”, International Journal of Advances in Scientiofic Research and Engineering, Vol. 4, No. 6, June 2018, pp. 122-126.

[13] Karthikeyan T., Thangaraju P., “PCA-NB Algorithms to Enhance the Predictive Accuracy”, International Journal of Engineering and Technology, Vol. 6, No. 1, 2014, pp. 381-387.

[14] Farid J., Ahmad A.S., “Building student’s performance cesissiion tree classifier using boosting algorithm”, Indonesian Journal of Electrical Engineering and Computer Science, Vol. 14, No. 3,2019, pp. 1298-1304.

[15] Babak M.A., Seyed K.S., Maryam M.M., “Support vector machine-based arrhythmia classification using reduced features of heart rate variablitiy singanal”, Ariticial Intelligence in Mechine (Elsevier), vol. 44, 2008, pp. 51-64.

[16] Humera S., Raniah Z., Kavitha G., “Prediction of Student Performance in Semester Exam Using a Naïve Bayes Classifier”, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, No. 10, October 201 5, pp. 9823-9829.

[17] Jollifie I.T, "Principle components analysis and factor analysis", Springer 1986.

[18] Phauk Sokkhey and Takeo Okazaki., "Comparative Study of Prediction Models on High School Student Performance in Mathematics", Journal of IEIE Transaction on Smart Processing and Computing, Vol. 8, No. 5, October 2019, pp. 394-404.

[19] Mohamed Z.G. A., Mustafa B. M., Lazim A., and Hamdan A. M., "The Factors Influence Students' Achievement in Mathematics: A Case for Libyan's Students ". Australian Journal of Basic and Applied Science, Vol. 17, N0. 9, 2012, pp. 1224-1230.

[20] Uysal S., "Factors affecting the Mathematics achievement of Turkish students in PISA 2012", Academic Journals, Vol. 10, June 2015, pp. 1670-1678.

Authors’ Profile

Phauk Sokkhey was born in Kompong Thom province, Cambodia. He received his bachelor degree in Mathematics from Royal University of Phnom Penh (RUPP), Cambodia, in 2010, and later received his Master degree in Applied Mathematics from Suranaree University of Technology (SUT), Thailand, in 2013. He was a lecturer of mathematics at Institute of Technology of Cambodia (ITC). Sokkhey is currently a PhD candidate at the University of the Ryukyus, Japan. His currently research are statistical causal relationship analysis, machine learning, educational data mining, and data science.

Takeo Okazaki took B.Sc., M.Sc. from Kyushu University in 1987 and 1989, respectively. He had been a research assistant at Kyushu University from 1989 to 1995. He earned his Ph.D. from University of the Ryukyus in 2014. He is currently a professor at the university of the Ryukyus. His research interests are statistical analysis, data analysis, genome informatics, tourism informatics, geographic information systems, and data science. He is a member of JSCS, IEICE, JSS, GISA, and BSJ Japan.

Proposal of a Sustainable Agile Model for Software Development

Oscar Antonio Alvarez Galán[1], José Luis Cendejas Valdéz[2], Heberto Ferreira Medina[3]
Gustavo A. Vanegas Contreras[4], Jesús Leonardo Soto Sumuano[5]
Engineering in Information Technologies, Universidad Tecnológica de Morelia, Morelia Michoacán, México[1, 2, 4]
Unit of Information Technologies and Communications, Instituto de Investigaciones en Ecosistemas y Sustentabilidad–UNAM
TECNM/I.T. de Morelia, Morelia Michoacán, México[3]
Centro Universitario de Ciencias Económico Administrativas (CUCEA). Universidad de Guadalajara (UdeG)[5]
Cuerpo Académico de Tecnologías de Información y Comunicación (UDG-CA-125)[5]

***Abstract*—Sustainability is even more important now than ever if we speak in the context of organizational growth, it is necessary that technological products, such as software developments, are certified as green- environmental friendly technology that would mean a competitive advantage for an organization that implements an agile methodology for software development that takes sustainability into account, giving the organization new ways to market their software products as environmentally friendly. This study proposes a model for agile software development, it has taken into account that software development must be based upon reusing old hardware, free non-privative software and code (open source), as well as virtualization of servers and machines, to create software that can be useful for over a decade, as a result, we expect a reduction of planned obsolescence in hardware, which means taking one step ahead to help solve the problem that the big amount of electronic waste (e-waste) means nowadays worldwide.**

***Keywords*—Sustainability; agile methodologies; software development; MDSIC; sustainability variables**

I. Introduction

Sustainability is more important every day if we talk about development processes at technological organizations. Today it is necessary that products and services are certified as green technology, there is data from Solving the e-waste problem (STEP) [1], that shows every person produces an estimated amount of 7 to 9 kilograms of electronic waste per year in Mexico, it is estimated that 42 billion tons of e-waste will be generated worldwide by 2016.

Technology companies generate their products and software and cloud services using traditional methodologies, these do not take into account economic sustainability or ecological sustainability, so devices that use this software have a predefined useful life for about 2 to 4 years approximately, it means they are fabricated with the idea that these devices, especially mobile ones will become obsolete within a short time, that way it increases the product demand [2], [3],[4], [5]. It has been shown that as time passes software becomes obsolete and can no longer update.

Companies that develop software require greater hardware resources so that their applications work in the best way, which generates a limited lifetime in the devices and the software that is built requires more resources for its optimal operation; thus, advancing the useful life and becoming technological waste.

The development of custom made software represents high costs for organizations and many of these projects finally do not meet its minimum requirements. Organizations with different business lines want to enhance their information processes with the help of software, made by so-called software factories. Those factories can help systematize and improve the processes of organizations, with the use of ecotechnologies costs will be reduced and using code that is sustainable software can be developed faster.

It is for this reason that mobile software development brings new challenges to the software industry, because mobility, interconnection, and simpler applications are growing in demand. In many APPs global delivery stores, a tendency can be observed towards buying-selling software known as "freemium", a free license with minimum features, which can be changed by paying to access the software full features, upgrades and improvements.

It is for this reason that this research proposes a methodological model so agile software can be developed and be useful for at least 10 years, that way it can make a difference in the efficiency of the essential process of organizations. This research solves many issues related to how green software can be developed, the proposed methodology on this research takes into account measuring sustainability indicators and attributes and shows how those are related to each step of the agile development of mobile software. It also proposes a full architecture that uses microservices and non-relational databases, so that sustainable software development can be achieved. Most modern companies can certify this kind of development as green software which represents a competitive advantage for organizations.

The rest of the work is organized as follows – the next section discusses briefly sustainability, ecotechnologies and previous methodologies based on the experience of this industry in Mexico. This experience allowed building a methodological process, that takes into account attributes and indicators so we can be sure that technological products developed using the proposed methodology are sustainable. Also, the paper presents the results obtained from the development of a software prototype and how we can measure the attributes and indicators on that software.

II. Literature Review

The Earth Charter is a declaration of fundamental principles for the construction of a global society in the XXI Century [6], which is fair, sustainable and peaceful. It inspires people to understand the concept of shared responsibility. The charter means hope and it is a call for action, it contains information about how to transition to sustainable lifestyles and sustainable human development. As a consequence of human activity, climate change is manifesting itself inevitably, bringing consequences for future generations [7]; which are mainly observed in the increase in temperature and the increase in sea levels, as well as the loss of species worldwide.

One of the most important entities in the care of the environment worldwide, the international organization, World Wide Fund for Nature (WWF) [8] [9], Living Planet report, presents a series of graphs based on different studies such as world population, carbon dioxide (CO.), fertilizer consumption, freshwater use, tropical forest loss, marine fish capture, which show the effects caused on the planet, due to the production of electronic waste of the human being, which are observed in the Fig. 1.

With the data in the previous graphs, technological proposals must have the attribute of sustainability as a basis [10], this concept refers to the efficient and rational management of resources so that it is possible to improve the well-being of the current population without compromise the quality of life of future generations. On [11] it is mentioned that sustainability sciences study this concept and propose solutions from a perspective that integrates various disciplines and the treatment of problems that prevent the transit to sustainable development.

Ideally, technology and software products should have an acceptable degree of resilience, that is, the ability to endure over time and the changes to which they are subjected, this concept is important to apply when talking about sustainability and technology, however, it is common for companies to focus more on economic benefit, giving the product a predetermined period of life.

Companies or organizations that do not take care of sustainability and resilience in their technological products fall into a practice known as planned obsolescence which is a production policy, it manages the useful life of an object being determined so that the object is then obsolete, useless or no function in a short time [5]. This industrial policy is potentially beneficial for the producer since it forces the consumer to acquire the same product several times, stimulating demand for it.

It is for the reasons described above that organizations and companies need to apply ecotechnologies in the development of their technological products, thus promoting sustainability and resilience, preventing them from becoming obsolete in a short time.

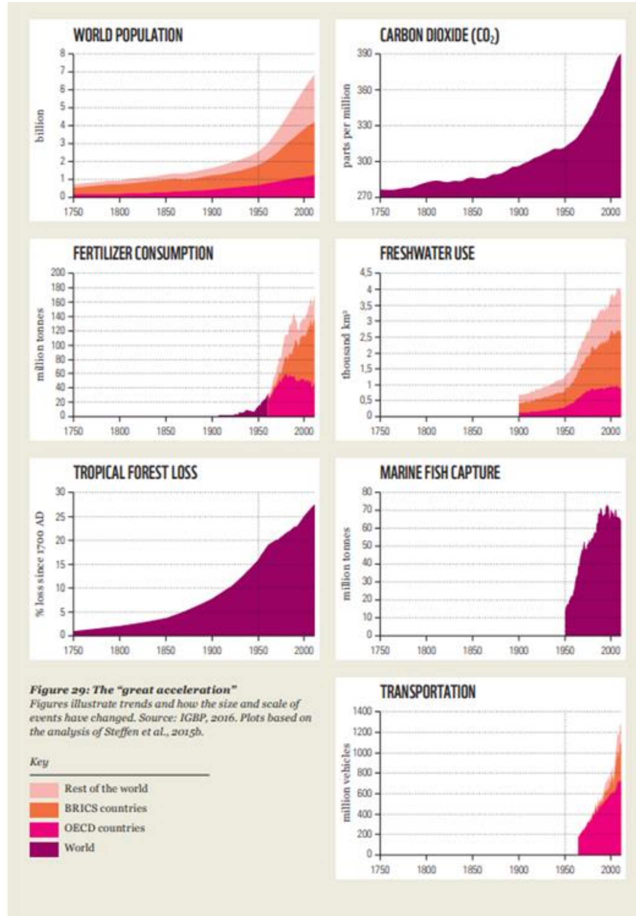


Fig. 1. Trends in Global Climate Change [8], [9].

It is mentioned that ecotechnologies "are the use of technological means for the management of ecosystems, based on a deep knowledge of the principles on which natural ecological systems are based and the transfer of this knowledge towards their management, so that the damage caused to the environment is minimized" [12] Among the most used green technologies today are solar panels, wind turbines, biogas chambers, energy-saving bulbs, solar heaters, dry toilets, etc.

It will be important that Information Technologies (IT) is seen in the organization as a great investment, since the use of technologies is a competitive differential for it, keeping IT up to date means a constant economic investment and Certifying technological products as sustainable can greatly help the return on investment. To achieve this denomination according to [13], [14] green technologies must meet certain criteria:

- Be accessible, especially to the poorest sectors of society.
- Be focused on local needs and contexts.
- Be friendly to the environment, promoting the efficient use of resources, recycling and reuse of products.
- Promote the use of local resources and their control.

- Generate employment in regional economies, especially in rural areas, from which the population has had to migrate due to a lack of opportunities.
- Preferably be produced on a small scale and in a decentralized manner.
- Be designed, adapted and disseminated through participatory processes, with dialogue between local knowledge and scientists.

It is for this reason that every implementation of technology in an organization should be aimed at cost reduction and that is why virtualization can be considered a green technology. It is mentioned on [15], [16] the need for departments or areas that are dedicated to generating the research, development and innovation process (R & D & I) is essential for most companies; this in order to enrich their processes. this department should ideally be in charge of applying a methodology that guarantees that the developed technological products have a useful life of at least 10 years and comply with the basic concepts of sustainability and resilience.

Among the existent models, on [15] it is proposed the Sustainable Technological Model (DESUSTEC), Which focuses on proposing a regulatory framework for the development of technology that is manufactured in software companies in Mexico. Being also sustainable, which is accompanied by good practices, is an alternative in the use of green technological infrastructure [15], [16].

As we can see in the Fig. 2 DESUSTEC consists of 6 levels, the first level consists of a software development cell made up of engineering students that will serve as a source of human resources for companies dedicated to software development, the second level refers to the use of Collaborative Integrated Model in Agile Software Development (MDSIC) which serves as a guide for the development of agile software, the third level refers to the use of ecotechnologies such as solar energy, virtualization and use of free software, the fourth level refers to the measurement of indicators and attributes of sustainability in the software, finally, the fifth level refers to the development of mobile software with a useful life of at least 10 years complying with the aforementioned indicators and attributes of sustainability, and the sixth level focuses on the reduction of electronic waste produced by mobile devices, servers, etc.

On the second level of DESUSTEC is contemplated MDSIC proposed in [17] which aims to develop quality software in small and medium software enterprises (SMEs), based on the alignment of information technologies with the essential process of the organization as well as being an agile model. MDSIC proposes five different levels that provide best practices for software development, in addition to being supported by the processes proposed by Project Management Institute (PMI).

MDSIC evaluates the quality of the software through a series of indicators that must be considered for optimal functioning. The first five elements that PMI mentions are integrated into the model: 1) Project integration, 2) scope, 3) time, 4) cost, 5) quality. This will allow the measurement of quality in a project attached to the standards proposed by PMI.

Fig. 3 details each of the MDSIC levels, which will help the software developed to be aligned with the essential process of the organization and its levels are:

- Level 0: Problem Detection
- Level 1: Analysis and Design,
- Level 2: Development,
- Level 3: Implementation,
- Level 4: Quality indicators.

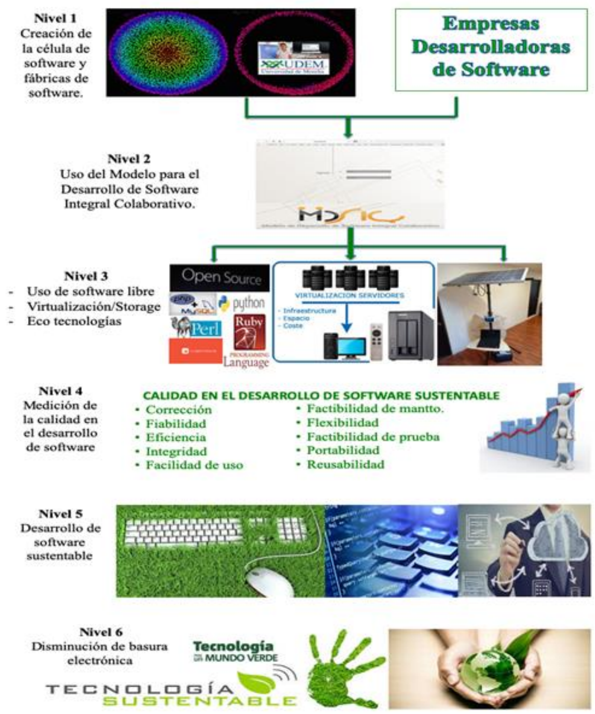


Fig. 2. DESUSTEC Model [15].

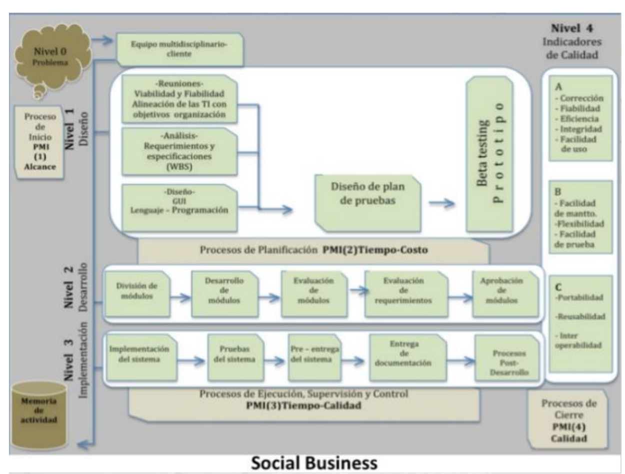


Fig. 3. MDSIC Model (Collaborative Integrated Model in Agile Software Development) [16].

Given the market trend towards the development of mobile technology [3] [18] proposes a version of MDSIC. Collaborative Integrated Model in Agile Software Development for Mobiles (MDSIC-M), it is a model based on MDSIC, for agile software development, which is oriented to rapid development and that meets the needs of small and medium companies, also proposes the documentation of the projects based on the model, through stories that facilitate the development of subsequent projects, as well as the use of their MDSIC 2.0 tool (software guide to apply the methodology in the development of a quality APP).

MDSIC-M elements are:

a) Establishment of requirements (tasks, deliverables and metrics),

b) Initial plan or WBS (Work Breakdown Structure),

c) Establish deliverables in the WBS,

d) Use of a scheduler for task control (scheduler),

e) Project time control by deliverable,

f) Risk control when allocating resources,

g) Cost control (establishment and metrics),

h) Guarantee the quality; KPIs (performance indicators) and the Scope (total time) of the project,

i) Use of a collaborative tool based on the scheduler,

j) Test and delivery plan (prototypes) and

k) Document lessons learned (database with project information).

The scheduler or planner is the core of MDSIC-M as it tracks and controls all the activities or tasks defined in the WBS or breakdown structure, specifies a list of activities with a hierarchy that allows building a complete map of all project activities, tasks, resources, times, costs and deliverables.

MDSIC-M proposes improvements to offer quality through KPIs; establishing roles, collaborative work teams, clearly defining deliverables, tools to use, standards to generate relevant information, control risks (costs) and finally ensure quality using techniques and standards. Fig. 4 shows MDSIC-M model.

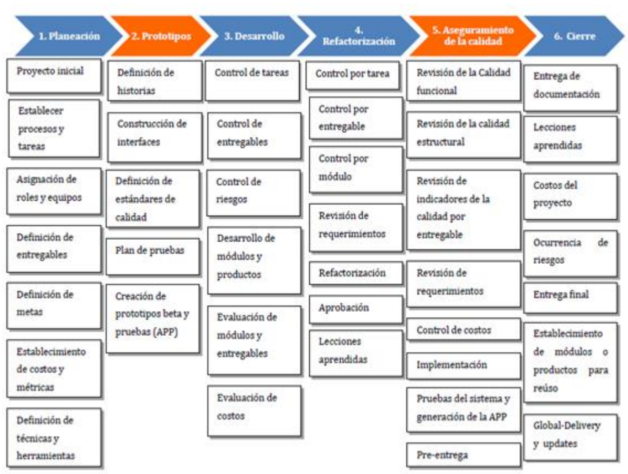


Fig. 4. MDSIC-M Model (Collaborative Integrated Model in Agile Software Development for Mobiles) [3].

Based on the aforementioned models and using the model proposed in this document, the degree of sustainability and resilience in the software will be measured through indicators and attributes, for this reason, the proposed model is also based on the Framework for the Evaluation of Natural Resource Management Systems Incorporating Sustainability Indicators (MESMIS) which incorporates sustainability indicators, MESMIS is a model that helps assess the sustainability of systems that manage natural resources, it defines a series of critical points or strengths and weaknesses for sustainability, which are related to three areas of sustainability. environmental, social and economic [19].

In this research, the proposal of a prototype of sustainable and agile mobile software was developed based on a methodological model that integrates the measurement of MESMIS sustainability indicators, together with the sustainability indicators obtained from the main correlations of this same study, with the objective that the proposed software prototype be sustainable and with a useful life of at least 10 years.

III. Methodology

It was determined that the nature of the present investigation is of type a) analytical, as indicated [20] since facts and information that are already available can be used and this material was analyzed to perform a critical evaluation, b) exploratory, which is used when the research problem or topic is poorly studied and has not been addressed before. That is when the literature review reveals that there are ideas vaguely related to the problem of study [21], c) quantitative, is based on the measurement of quantities, it is applicable to phenomena that can be expressed in terms of quantity. d) non-experimental, it is done without deliberately manipulating variables, no situation is constructed, but situations are observed, it is not intentionally provoked, there is no control over the variables [20]. e) documentary, which defines [21] how to detect, obtain and consult bibliography and other materials that participate in other knowledge and / or information collected moderately in a selective way so that they can be useful to identify studies.

Table 1 shows how the research for the construction of sustainable software was carried out in 5 main stages which include 1) Characterization of sustainable software, where a background check was made regarding sustainability and development of software with green technology, 2) independent and dependent variables which will be measured, as well as to identify what were the sustainability variables not measured by MDSIC-M (Collaborative Integrated Model in Agile Software Development for Mobiles) [3], 3) Application of a survey to professionals of sustainability sciences and software developers, where the variables identified above were measured, 4) The proposal of a model, based on DESUSTEC in which the most important correlations obtained in the survey are analyzed and based on them defined the most important variables to be measured during the process of software development with MDSIC-M, 5) The creation of a software prototype for the measurement of these indicators, following the methodology, to guarantee the quality in the development.

TABLE. I. METHODOLOGICAL MODEL

1.-Characterization of sustainable software development
Characteristics of sustainable software development or identification of variables. Cross-sectional analysis of existing models and methodology.
2.-Objectives of the methodology
According to the objectives, establish the dependent and independent variables to analyze. Establish variables not measured using MDSIC.
3.-Survey of professionals in sustainability sciences and sustainable development
Analyze the acceptance and use of the elements of the DESUSTEC model in Mexico. Establish the population of the survey according to a confidence interval. Analyze the consistency of the surveys applied using Cronbach's alpha. Establish correlations between the census variables (Pearson).
4.-Proposal of the methodology
Analyze the correlations and establish the importance of each of the census variables. Establish the most important sustainability variables and their alignment with a development methodology. Propose a methodology, using standards and best practices.
5.-Methodology tests
Design the elements of the methodology and rely on the development of software to simplify its use. Perform a quasi-experiment to verify the feasibility of use. Document results.
6.-Technology Transfer
Perform a quasi-experiment to verify the feasibility of use. Document results.

a. Table 1.- Methodological model.

Teachers of the state of Michoacán were selected for this research, in [22] there are 8,848 teachers of higher education (population of professionals, whose lines of research focus on IT and the environment). From the data obtained previously and with a 95% confidence interval and a margin of error of 5% [20], according to the method of finite populations, the sample size was obtained, (1) was applied and the replacement of the data is shown in Table 2.

$$n = \frac{z^2.p.q.N}{(e^2(N-1))+(z^2.p.q)} \quad (1)$$

Where:

N = The size of the population to be registered (Higher education teachers in Michoacán).

z = Area under the normal curve with confidence interval: 95% q = 1 - p = 0.5 Proportion of the sample for a maximum n with greater precision.

p =% of times the phenomenon occurs in the population.

e = margin of error of the calculated sample (5%).

TABLE. II. REPLACEMENT OF DATA

Confidence level 95%	p= 95%	e = 5%
z= 1.96	(1-p)= 5%	N = 8848

n=72

Once it was carried out, it was analyzed, where the alpha Cronbach coefficient [23] was obtained, with a value of 0.737 which indicates an acceptable internal consistency in the survey results. The most significant Pearson correlations obtained are those with values> = 0.5, which are detailed in Table 3.

TABLE. III. MAIN CORRELATIONS

1.-Direct contact with users and constant feedback.
2.-Correct analysis of the problem to be solved and requirements.
3.-Free Software Implementation
4.-Use of eco technologies for saving electricity consumption.
5.-Reused materials and components (hardware)
6.-Efficient use of natural resources
7.-That the software on the devices is kept updated.
8.-Good management of risk management in planning.
9.-Guarantee the useful life of the software for at least 10 years of operation.
10.-Compatibility even on older machines or different operating systems.
11.-Support, connection with other applications, maintenance and speed, stand out as quality indicators for the software.
12.-Ease of use, utility, reliability, performance, and compatibility should be a priority to develop software.
13.-Proper planning and management of the same project must be carried out supported by tools to facilitate the tasks.
14.-Defining clear roles and objectives, each member of the development team must know what to do.
15.-Management of times previously established in the planning, to obtain software with quality, this must be developed in the times dictated from the beginning.
16.-Use of virtualization technologies that will allow us to use virtual versions of devices, meaning savings in energy consumption and hardware.

IV. DISCUSSION AND ANALYSIS

The proposed methodological model based on MDSIC-M is illustrated in Fig. 5.

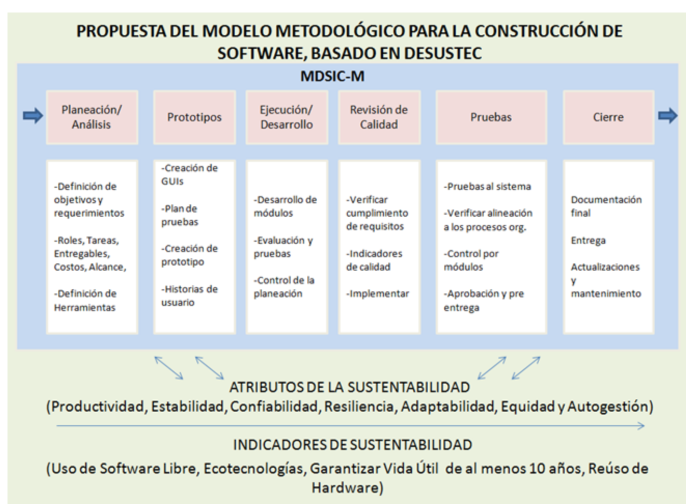


Fig. 5. Proposal of the Methodological Model for Software Construction, based on DESUSTEC.

Table 4 describes how each of the sustainability attributes based on MESMIS will be evaluated in the software and how it relates to the sustainability indicators obtained in the main correlations of this study.

TABLE. IV. ATTRIBUTES AND INDICATORS

Atributes	Equation	Related indicators	Equation
Productivity	The number of modules developed / Development time.	Life of at least 10 years of the software	10-year return on investment
Stability	The number of errors detected in tests / The Number of modules developed.	Use of free software	Total number of free software used / Total software
		Life of at least 10 years of the software	10-year return on investment
Reliability	Amount of money invested in development / Development time	Use of free software	Total number of free software used / Total software
Resilience	The number of functional modules / Number of years in operation	Use of free software	Total number of free software used / Total software
		Ecotechnologies	Number of applied ecotechnologies / Total applied technologies
		Life of at least 10 years of the software	10-year return on investment
		Hardware reuse	The number of reused hardware parts / Total hardware used
Adaptability	Frequency of use of ecotechnologies and reuse of software / number of years in operation	Ecotechnologies	Number of applied ecotechnologies / Total applied technologies
		Life of at least 10 years of the software	10-year return on investment
		Hardware reuse	Number of reused hardware parts / Total hardware used
Equity	The number of users / The number of free software used	Use of free software	Total number of free software used / Total software
		Ecotechnologies	Number of applied ecotechnologies / Total applied technologies
		Hardware reuse	Number of reused hardware parts / Total hardware used
Self-management	Total current cost of software development / Initial cost of software development	Life of at least 10 years of the software	10-year return on investment

It is proposed as good practices for software construction to follow sustainability indicators, which allow to generate a uniform architecture using [24], [25], [26] for the layers of the applications, separating them into microservices that are implemented in an API as a communication protocol between the client and the server, facilitating its maintenance, scalability, and compatibility in the software through these sustainable indicators. Table 5 shows the proposed layers.

The prototypes developed in future works based on this model will measure attributes and indicators using the equations already shown, and it will be known the level of sustainability that each software development has, so companies keep using it every time they develop new software [27], [28].

An example of the application of the model proposed was built as a prototype that shows a CRUD (Create, Read, Update, Delete) for a general non-relational database, where the use of the code is shown through generic libraries in Swift. For this application, the indicators and attributes used in the proposed model were used.

V. RESULTS

The software prototype built using this model as a base uses functions that allow receiving any number of parameters in an array which is part of the data sent by the client in a request to the server to insert, update, delete or request data, is generated in this way so that the code is reusable for any context, the prototype allows the names of the keys and values to be assigned to convenience, open-source programming language swift 3 was used for this prototype as well as [29], [30], [31] the uniform architecture for sustainable software development proposed on Table 5.

TABLE. V. UNIFORM ARCHITECTURE FOR SUSTAINABLE SOFTWARE DEVELOPMENT

API (Application) • Interaction with client / server by HTTP protocol allows interaction between both	Interface layer at the protocol level. In the prototype developed, the API interacts with the server through HTTP routes, through which REST requests can be sent
Frontend (View) • It is the interface between the user and the system, implements interaction techniques and methods. Construction tools with good practices are taken into account.	Client side layer. It is the graphic interface of the prototype developed
Middleware (intermediary) • It is an intermediate software layer between the operating system and communication networks (services), below the user's applications	The intermediate layer of communication between client and server
Backend (base software) • It is the information access layer that defines the server-side database	Server side layer. A server that handles the data in the prototype, for this development the NOSQL server Apache Couch DB (Apache, 2016) was used.

Based on [25], [26].

Fig. 6 shows some examples of the functions used within the software prototype:

```
    func insert() {
        items.append("Item \(items.count + 1)")

        let insertionIndexPath = IndexPath(row: items.count - 1, section: 0)

        tableView.insertRows(at: [insertionIndexPath], with: .automatic)
    }

    override func tableView(_ tableView: UITableView, numberOfRowsInSection sec-
tion: Int) -> Int {
        print("NUMERO DE ITEMS=\(items.count)")
        return items.count

    }

    override func tableView(_ tableView: UITableView, cellForRowAt indexPath: Index-
Path) -> UITableViewCell {
        let myCell = tableView.dequeueReusableCell(withIdentifier: "cellId", for: index-
Path) as! MyCell
        myCell.nameTextField.text = items[indexPath.row]
        print("VALORES DE ITEM=\(items)")
        myCell.myTableViewController = self

        return myCell
    }

    override func tableView(_ tableView: UITableView, viewForHeaderInSection sec-
tion: Int) -> UIView? {
        return tableView.dequeueReusableHeaderFooterView(withIdentifier: "headerId")
    }
```

Fig. 6. Software Prototype Functions, based on the Proposed Methodological Model using Swift Code [32].

In Figure 7 and 8 two interfaces of the software prototype developed during the present investigation are observed.

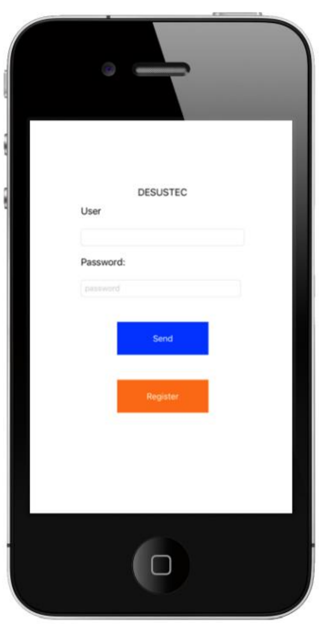


Fig. 7. Software prototype login screen, based on the proposed methodological model.

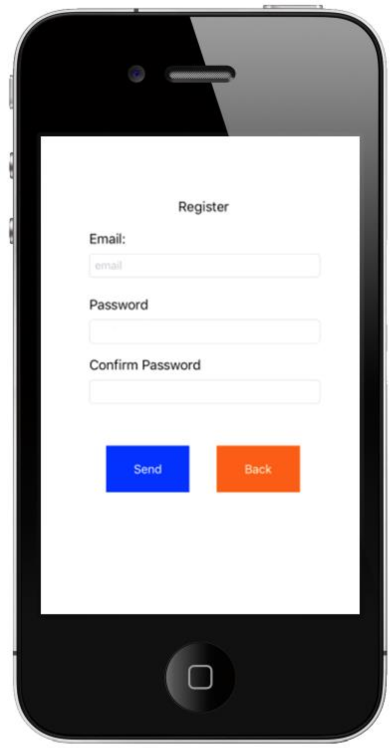


Fig. 8. Software Prototype, Register Screen.

VI. Conclusion

According to the data and calculations of sustainability indicators obtained, it can be seen that there is still work to be done so that the model improves, along with how sustainability is ensured in software development. Since the measurement of indicators in the proposal depends on the number of years elapsed and how productive the software remains after these; understanding that sustainability in software is just being explored since it is a new issue and there is still much more to understand and improve in this field.

According to the revised agile methodologies for the model, none takes sustainability into account, so this represents an advantage over others in its application. Software developed under the proposed model may be marketed as green technology. Saving energy is encouraged, implementing ecotechnologies such as virtualization for the use of microservices which represents a competitive advantage for organizations, as well as the reduction in production costs and time. Since one of the main pillars of the proposal is the reuse in code to speed up the creation of the software. It is intended to generate in a future project a uniform infrastructure that allows sustainably to offer the creation of microservices [28], offering SaaS (software as a service), [29], [30], a uniform architecture that encompasses all the technologies that today are scattered, helping users to develop software quickly with quality. The proposed model may be a sustainability framework applicable to technological projects, even in other areas of knowledge. Today the legal framework, begins to request in several countries that companies contemplate ethically and socially to the sustainability in its processes for the generation of products and / or services.

ACKNOWLEDGMENT

We are grateful for the support of the Ecosystem Research Institute, to the Technological University of Morelia. Especially to Eng. Atzimba Graciela López Maldonado., MTI. Alberto Valencia García And Hugo A. Zavala V. his comments and support in the analyzes. To Eng. Diego Cabrer F. for his help in the application of the model.

REFERENCES

[1] STEP, «Solving the E-wast problem» 2016. ISSN:2071-3576.

[2] Asociación Mexicana de Internet, «Dispositivos Móviles: Hábitos del consumidor Mexicano,» 2014. [Online].

[3] H. Ferreira, «Metodología MDSIC-M para el desarrollo de aplicaciones móviles en empresas Pymes,» 2015. [Online].

[4] K. Brigden, I. Labunska, D. Santillo, & M. Allsopp, Greenpeace, «Recycling of electronic wastes in China and India: Workplace and environmental contamination,» 2005. [Online].

[5] P. Ramírez, «Obsolescencia tecnológica programada,» Facultad de ciencias y tecnología, Universidad Católica Sede Regional Asunción, 2012.

[6] Iniciativa Carta de la Tierra, «Carta de la Tierra,» 2016. [Online].

[7] G. Gaidajis, K. Angelakoglou y D. Aktsoglou, «E-waste: Environmental Problems and Current Management,» 2010. [Online].

[8] World Wide Fund for Nature, «Planeta Vivo Informe 2016» 2016. [Online]. ISBN: 978-2-940529-48-3.

[9] Elsevier, «Sustainability Science in a Global Landscape,» Elsevier's Analytical Services, 2015.

[10] G. Merino, R. Tovar « Gasto, inversión y financiamiento para el desarrollo sostenible en México,» Proyecto CEPAL/PNUD RLA/01/001 "Financiamiento para el Desarrollo Ambientalmente Sostenible en América Latina y el Caribe", 2002. ISBN: 9213220995.

[11] T. Ortiz, A. Quiroz, C. Migoya, « Evaluación de Sustentabilidad Para el Manejo de Sistemas Naturales Manual Para el Técnico Comunitario » Basado en el Marco de Evaluación de Sustentabilidad: MESMIS 2016. [Online].

[12] L. Gonzaga Gutiérrez, «Perspectivas de la biotecnología en las ecotecnologías,» Scientia et Technica, p. 451, 2006.

[13] J. A. Ortiz, O. Masera y A. F. Fuentes, La Ecotecnología en México, México D.F.: Imagia, 2014.

[14] C. Ferrete Sarria, «La eco-tecnología y la ciencia ecológica como alternativas. una larga historia de olvidos,» 1999. ISSN 1132-8177.

[15] J. L. Cendejas Valdez, H. Ferreira Medina, G. A. Vanegas Contreras y M. A. Acuña Lopez, « El modelo sustentable tecnológico (DESUSTEC) como herramienta en el desarrollo de software para la reducción de basura electrónica » Journal of Technology and Innovation, p. 5-15, 2019.

[16] T. Mackey, J. K. Benedict « XenServer Administration Handbook: Practical Recipes for Successful Deployments,» 2016. ISBN 13: 9781491935439.

[17] J. L. Cendejas Valdéz, C. A. Vega Lebrún, A. Careta Isordia, O. Gutiérrez Sánchez y H. Ferreira Medina, «Diseño del modelo integral colaborativo para el desarrollo ágil de software en las empresas de la zona centro-occidente en México,» Revista Electrónica Nova Scientia, p. 133 – 148, 2014.

[18] H. Ferreira, J. L. Cendejas Valdez, H. Ferreira Medina, G. A. Vanegas Contreras y G. Cortes Morales, A. H. Ginori González «Collaborative Integrated Model in Agile Software Development (MDSIC/MDSIC–M)-Case Study and Practical Advice » 2019. [Online].

[19] M. Astier, O. Masera, L. García, C. González, S. López-Ridaura, T. Ortiz y Y. Galván-Miyoshi, «MESMIS - Marco para la Evaluación de Sistemas de Manejo de recursos naturales incorporando Indicadores de Sustentabilidad,» 2012.

[20] C. Kothari, Research Methodology : Methods And Techniques, New Delhi: New Age International, 2019. ISBN-10: 9386649225.

[21] R. Hernández, C. Fernández y L. Baptista, Metodología de la investigación, México: McGraw Hill, 2000.

[22] Secretaría de Educación Pública Dirección General de Planeación, Programación y Estadística Educativa « Estadística educativa Michoacán de Ocampo Ciclo escolar 2018-2019, » 2018.

[23] L. J. Cronbach, «Coefficient alpha and the internal structure of tests,» , pp. 297–334, 1951.

[24] Apache, «Apache HTTP Server Implementation Notes,» 2016. [Online]. Available: https://iiif.io/api/annex/notes/apache/#table-of-contents. [Accessed: 06 Apr 2017].

[25] K. Ayala, H. Ferreira, R. Ferreira, A. Nuñez y J. Ruíz, «Caracterizacion de WebApp para el desarrollo web - Arquitectura uniforme,» 2017.

[26] W3C, «Web Services Architecture,» 2004. [Online]. Available: https://www.w3.org/TR/ws-arch/wsa.pdf. [Accessed: 10 Sep 2016].

[27] P. R. Fillottrani, «Calidad en el Desarrollo de Software - Modelos de calidad de software,» 2007. [Online].

[28] G. Sunkel, D. Trucco, A. Espejo, La integración de las tecnologías digitales en las escuelas de América Latina y el Caribe. Una mirada multidimensional, Santiago de Chile: Naciones Unidas Cepal, 2014. ISBN 978-92-1-121851-0.

[29] M. Fowler, «Microservices: a definition of this new architectural term,» 2014.

[30] Microsoft, D. Sprott y L. Wilkes, «Understanding Service-Oriented Architecture,» 2004.

[31] M. D. Gonzalez Rio, Tecnologías de Virtualización, Createspace, 2014.

[32] Apple Inc. (2014). The Swift Programming Language. Apple Inc.

Pathological Worrying and Artificial Neural Networks

Carlos Pelta
Dept. of Experimental Psychology,
Cognitive Processes and Speech Therapy, UCM Madrid, Spain

***Abstract*—Worrying is a cognitive process that focuses on potential future negative events, where the outcome is often uncertain. Worries can arise in chains with one worry leading to another, often without solution. This may give rise to an uncontrollable worrying that may be associated with psychiatric disorders such as anxiety and depression. The generation of progressively more negative chains of worries can lead to a catastrophic phenomenon of pathological worrying. In this article we show that catastrophic worrying can be simulated by using a cascade-correlation algorithm for artificial neural networks.**

***Keywords*—Pathological worrying; artificial neural networks; cascade-correlation algorithm**

I. Introduction

Worrying is a daily experience for most people but this does not mean that it is a simple process. Following a definition by Barlow [1], it is a cognitive process, typically verbal, that focuses on potential future negative events, where the outcome is often uncertain. For Wells and his metacognitive model [2], on the other hand, worrying is connected to metacognitive beliefs (both positive and negative) that imply a recurrent style of thinking linked to people's conscious strategies to deal with threats. Worries are interpreted as cognitions, in the line of cognitive behavior therapy or as metacognitive beliefs [3]. Worries can arise in chains with one worry leading to another, often without solution. This may give rise to an uncontrollable worrying that may be associated with psychiatric disorders such as anxiety and depression [4]. It has been suggested that non-pathological worrying and pathological worrying do not differ in content but in terms of prevalence and severity. Ruscio, Borkovec and Ruscio [5] have examined whether the difference between pathological worrying and normal worrying is one of a kind of degree. According to them, pathological and normal worrying take care of similar topics, but pathological worrying concerns to more issues, is more durable and often is uncontrollable.

The generation of progressively more negative chains of worries can lead to a catastrophic phenomenon [6]. When we think catastrophically, we imagine the imminence of some disaster. Following [7], catastrophism is based on an "overestimation" of the likelihood of a bad result, as well as an "undervaluation" of our ability to deal with the situation correctly. There is a catastrophic level of worrying because stopping the process becomes incompatible with beliefs about the need to worrying and with beliefs about uncontrollability [8]. A non-adaptive pattern of thought control is developed. This pattern prevents the person from discovering that the worrying process is controllable. Individuals use self-regulation strategies that are counterproductive to reduce worrying.

Using a procedure based on the technique of cognitive therapy of decatastrophizing [9], Vasey and Borkovec [6], found that chronic worriers generated more catastrophic steps than nonworriers and reported a significant increase in discomfort. The worriers spend more time ruminating that nonworriers.

In this study we propose to demonstrate how the use of artificial neural networks can simulate in a very reliable way pathological worry in a sample that mixes worriers and nonworriers subjects. For this, we have used the `Penn State Worry Questionnaire´ (PSWQ) in its Spanish version and a catastrophization questionnaire based on the catastrophization interview technique [10]. The results obtained have served to design an artificial neural network that applies a cascade-correlation algorithm for classifying pathological patients which takes more steps in the catastrophization interview. We show that artificial neural networks in cascade fit perfectly with the process of catastrophization typical of pathological worrying.

The use of artificial neural networks to model neurological and psychiatric disorders is not new. Witnesses to this are classics in the field such as [11] and [12]. They discuss the simulation of neurological disorders ranging from Alzheimer´s disease [13] to focal epilepsy [14] and neuropsychiatric disorders ranging from depression [15] to the study of the nature of schizophrenia and autism [16]. In fact, the new field of study of computational psychiatry [17] seems to be a very promising investigation program.

Predictive studies on general anxiety disorder (GAD) are also abundant [18]. However, our study focuses on the prediction of pathological worrying, a type of disorder that is central to the diagnosis of GAD or depression. The psychological disorder of pathological worrying is at the base of much of the psychological suffering and, until now, had not been modeled. Precisely, to this end our study aims to contribute.

We will begin with the selection of the experimental sample through the application of the PSWQ and the catastrophization questionnaire to then propose that what we believe is the best way to model a phenomenon such as the pathological worrying. It is to apply a cascade-correlation algorithm, because the cascading structure of the steps conforms to the computational structure of an artificial neural network in cascade. Next we will propose the architecture of our artificial neural network that simulates the phenomenon and we will carry out the evaluation of the training of the network to conclude and in some conclusions and possible applications of the model in the future.

II. Method

A. Procedure

We have applied the PSWQ [19] to a sample of *N* = 86 (*M* = 21.6, *SD* = 2.5) of high school students night shift (54% of female students and 46 % of male students). We have employed its Spanish version [20] which validates the original questionnaire and demonstrates high internal consistency in the case of both clinical and non-clinical criterion groups.

The PSWQ is a questionnaire that assesses worrying as an uncontrollable phenomenon. It is made up of 16 items, which asks participants to rate each item on a 5-point scale (“not at all typical” to “very typical”). Scores range from 16-80 and a higher score represents a greater degree of pathological worry.

Besides we have used an adaptation of the catastrophic interview procedure, focused on the topic “getting pass to the second year course” that gives access to the secondary school diploma. Subjects had to indicate from 0 to 10 their degree of discomfort if they could not pass the course and the probability that it could happen (also on a scale of 0 to 10). Table I shows the catastrophizing sequences generated by a pathological worrier.

B. Results

The results concerning to the application of the Spanish version of the PSWQ were that 57.1% of subjects were nonworriers and introduced an average of 8.01 steps in the catastrophization procedure. The average factor of nonworrying was 0.78. In contrast, 42.9% showed a pathological worrying index in the PSWQ, validating an average of 14.1 steps and the average factor of worrying was 0.83. There was a significant difference between the number of steps between the PW (pathological worriers) group and the NW (nonworriers) group [$F(1,84) = 14.075$, $p < .05$]. The average degree of discomfort among the pathological worriers was 8.85 on a scale of 10. On the other hand, the average degree of discomfort among the nonworriers was 6.46. With regard to the average probability of occurrence of the aforementioned events, among the pathological worriers it was 0.81 and among the nonworriers it was 0.58. There were significant differences in degree of worrying between women and men [$F(1,84) = 14.112$, $p < .05$].

C. ANN Model Design

In this study we propose to demonstrate how the use of artificial neural networks can predict in a very reliable way the distinction between pathological worriers and nonworriers in a sample of students. We have designed an artificial neural network that applies a cascade-correlation algorithm that fits very well with the process of catastrophization, typical of pathological worriers. The cascade-correlation architecture [21] adds hidden neurons one by one in the network. In the process of adding hidden neurons to the network, each new neuron receives a synaptic connection from each of the input neurons and also from the hidden neurons that precede it. After adding each new hidden neuron, the synaptic weights of its inputs are frozen, while the weights of its outputs are repeatedly trained. This process continues until a desired performance is achieved [22]. It is evident that the cascade architecture allows adding each hidden neuron at a time and only the new weights are updated; the incremental or constructive learning allows to create the new hidden units, where for each new hidden neuron, the algorithm maximizes the magnitude of the correlation between the new hidden neuron and the residual network error, that is, hidden neurons are added trying reduce the network error until its performance is satisfactory.

We have created a cascade backpropagation artificial neural network that presents two input neurons (one that encodes nonworriers-NWI- and the other encodes pathological worriers-PWI-), a layer hidden (H1) of 8 neurons (C1-C8) and 7 hidden (H2-H8) cascading neurons (CPW1-CPW7) that are added to the successive hidden layers culminating in an output that encodes the pathological worriers' output (PWO). The output that encodes nonworriers (NWO) is connected to the 8 cascading neurons (C1-C8) of the first hidden layer (H1). Actually, with the seven hidden (H2-H8) cascading neurons (CPW1-CPW7), we have wanted to reflect the steps of more than, in the questionnaire, pathological worries had added. As shown in Fig. 1, our network has a pool of 15 candidates forming the layers of hidden neurons, two inputs (one exemplifying the pathological worries and the other exemplifying the nonworriers) and two outputs (to reflect the results of the pathological worrying and non worrying, respectively).

A sample of one pattern visualization showing the activation of PWI, PWO and of the hidden neurons is presented in Fig. 2.

TABLE. I. CATASTROPHIC SEQUENCES

Steps	**If I don´t get pass to the second year course**		
	Catastrophizing sequence	***Discomfort***	***Likelihood***
1	I won´t live up to my expectations.	9	9
2	It would be harder to study what I don´t like.	10	9
3	It would be harder to get a good job for me.	10	8
4	My salary would be lower.	9	9
5	It would demotivate me a lot.	10	9
6	I´d lose my self-confidence.	8	8
7	My loss of confidence would affect other areas of my life.	9	8
8	I´d become very anxious.	9	9
9	I´d feel like I wouldn´t have any control over my life.	10	8
10	I´d feel mentally bad.	10	8
11	I´d be very unhappy.	10	7
12	I´d like to die.	10	6

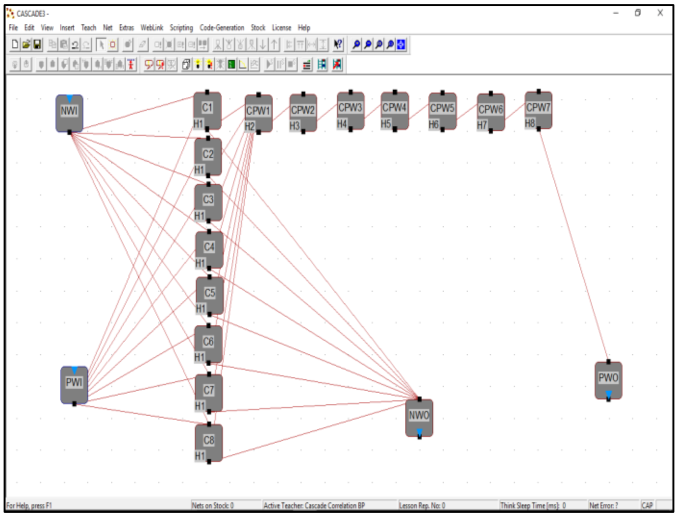


Fig. 1. The Proposed ANN Architecture.

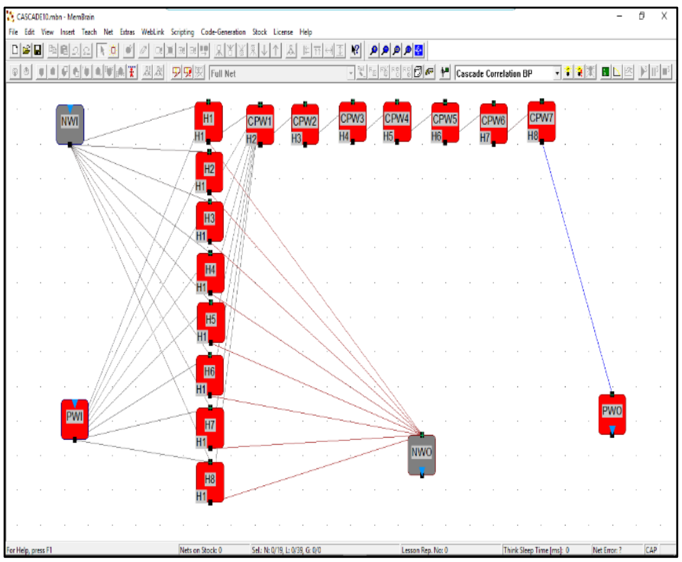


Fig. 2. Sample of Pattern Visualization.

III. ANN Evaluation Results

MemBrain (version V03.08.01.00), a framework designed to develop artificial neural networks was used in training our ANN model. It contains (see [23], [24], [25]) an implementation for most of the learning algorithms including cascade-correlation using backpropagation. We have used a Sigmoid (logistic) function as transfer function with values between -0.5 and 0.5. As observed in Fig. 3, learning rate was 0.25 and target net error was 0.01.

The learning rate parameter determines the size of the weights adjustment each time the weights are updated during training [26].

The initial weights have been matched to the probabilities with which the subjects have token the successive steps in the catastrophization questionnaire (see Table II).

Error measure has been then computed to assess the neural network´s accuracy. Mean Squared Error (MSE) has been used yielding the smallest value of MSE equal to 0.00986309276170759 for nonworriers and the smallest value of MSE equal to 0.009746931044799933 for pathological worriers (see Table III).

Fig. 4 and 5 show the Net Error Graph for nonworriers and pathological worriers.

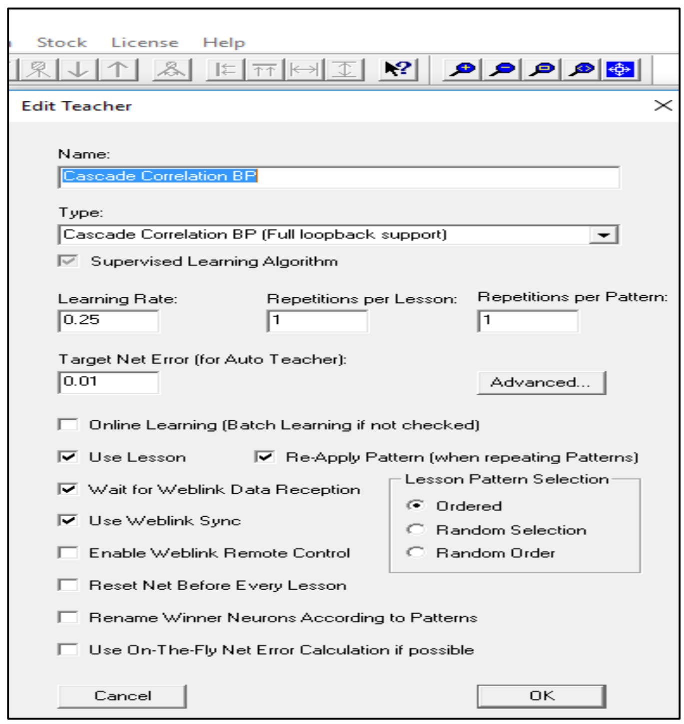


Fig. 3. View of "Edit Teacher" Tool.

TABLE. II. Initial Weights C1 C2 C3 C4 C5 C6 C7 C8 CPW1 CPW2 CPW3 CPW4 CPW5 CPW6 CPW7 *C* Candidate Neurons

Input layers	PWI	0.57	0.60	0.68	0.71	0.75	0.78	0.82	0.89	0.83	0.87	0.94	0.98	0.72	0.75	0.65
Input layers	NWI	0.13	0.15	0.22	0.23	0.27	0.28	0.34	0.37							
Output layers	PWO									0.84	0.89	0.95	0.99	0.74	0.76	0.71
Output layers	NWO	0.15	0.16	0.24	0.27	0.31	0.34	0.36	0.41							

TABLE. III. MEAN SQUARED ERROR

ANN	***Pathology***	***MSE***
Cascade	Nonworriers	0.00986309276170759
Cascade	Pathological worriers	0.00974693104479933

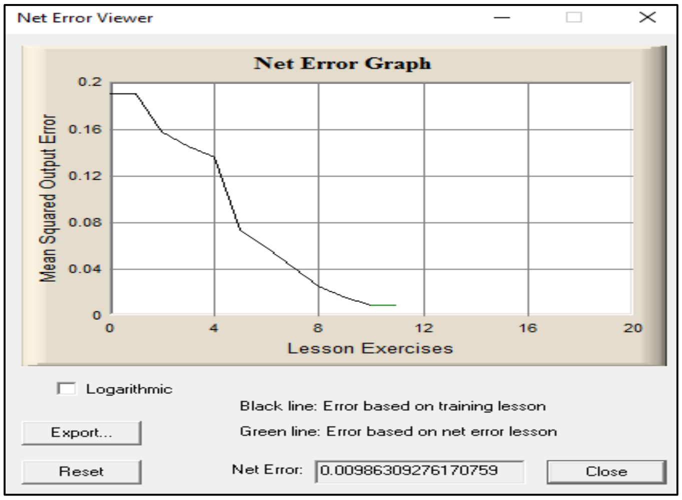


Fig. 4. Net Error Graph for Nonworriers.

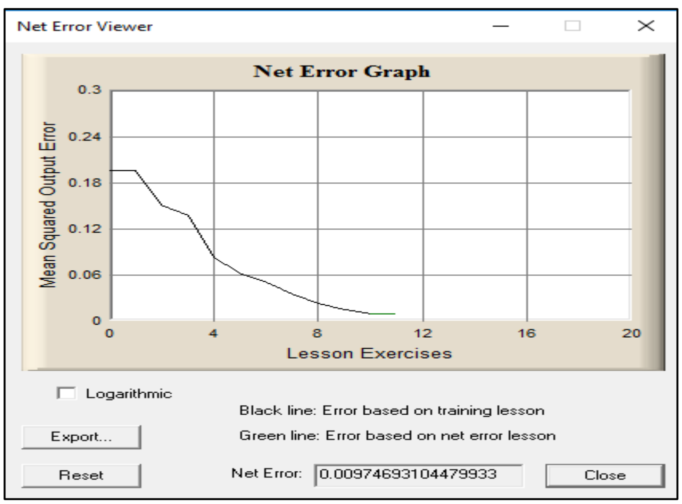


Fig. 5. Net Error Graph for Pathological Worriers.

IV. CONCLUSIONS

This study has attempted to build a suitable ANN model for modeling the phenomenon of worrying distinguishing between nonworriers and pathological worriers. The type of artificial neural network used has been a cascade-correlation network with backpropagation. The backpropagation algorithm modifies network weights to minimize the mean squared error between the desired and the actual outputs of the network. But cascade-correlation enables to create a constructivist topology [27] of hidden layers that reflects the catastrophization process typical of pathological disorders like worrying, rumination or chronic anxiety. Cascade-correlation seems more in tune with the simulation of events involving intense emotional cycles [28]. These cycles are self-amplifying positive feedback loops of strong negative emotions. Cascading networks grow in a similar manner to the process of catastrophization in psychiatric disorders and mimic its sequential growing in steps.

There were some limitations to note. First, a relatively small sample size and low reports of some behaviors. Further, measures of worrying used here need additional validation. It would be necessary to determine if worrying is actually difficult to control and to examine specific aspects of negative emotion and rumination that may be involved in pathological worrying.

A future task should be to simulate, by using cascade-correlation artificial neural networks, mental pathologies such as pathological rumination [29] or borderline personality disorder [30].

REFERENCES

[1] D. Barlow, Anxiety and its disorders: The nature and treatment of anxiety and panic, 2nd ed. New York: Guilford Press, 2004.

[2] A. Wells, "The metacognitive model of worry and generalised anxiety disorder," in Worry and its psychological disorders: Theory, assessment and treatment, G. T. L. Davey and A. Wells (Eds.) Chichester: John Wiley&Sons, 2006, pp.179–199.

[3] A. Beck, Depression: clinical, experimental and theoretical aspects. New York: Harper and Row, 1967.

[4] M.G. Newman, S.J. Llera, T.M. Erickson, A. Przeworski, and L. G. Castonguay, "Worry and generalized anxiety disorder: a review and theoretical synthesis of evidence on nature, etiology, mechanisms, and treatment," Annu. Rev. Clin. Psychol., vol. 9, pp. 275–297, March 2013. https://doi.org/: 10.1146/annurev-clinpsy-050212-185544.

[5] A.M. Ruscio, T.D. Borkovec,, and J. Ruscio, "A taxometric investigation of the latent structure of worry. Abnormal Psychology, 110(3), pp. 413-422, 2001. http:// dx.doi.org/10.1037//0021-843X.110.3.413.

[6] M.W. Vasey and T.D. Borkovec, "A catastrophizing assessment of worrisome thoughts". Cogn. Ther. Res., 16(5), pp. 505-520, 1992. https://doi.org/10.1007/BF01175138.

[7] E. Bourne and L. Garano, Coping with anxiety: ten simple ways to relieve anxiety, fear, and worry, 2nd ed., Oakland: New Harbinger Publications, 2016.

[8] A. Wells, "Metacognition and worry: a cognitive model of generalised anxiety disorder", Behavioural and Cognitive Psychotherapy, 23, pp. 301-320, 1995.

[9] P.C. Kendall and R.E. Ingram, "The future for cognitive assessment of anxiety: let´s get specific," in Anxiety and stress disorders: Cognitive-behavioral assessment and treatment, L. Michaelson and L.M. Ascher (Eds.) New York: Guilford Press, 1987, pp. 89-104.

[10] G.C.L. Davey, "The catastrophising interview procedure," in Worry and its psychological disorders: Theory, assessment and treatment, G. T. L. Davey and A. Wells (Eds.) Chichester: John Wiley&Sons, 2006, pp.157–176.

[11] J.A. Reggia, E. Ruppin and R. Sloan Berndt (Eds.), Neural modeling of brain and cognitive disorders. Singapore: World Scientific, 1996.

[12] D.J. Stein and J. Ludik (Eds.), Neural networks and psychopathology. Cambridge: CUP, 1998.

[13] E. Ruppin, D. Horn, N. Levy, and J.A. Reggia, "Computational studies of synaptic alterations in Alzheimer´s disease", in Neural modeling of brain and cognitive disorders, J.A Reggia, E. Ruppin and R. Sloan Berndt (Eds.) Singapore: World Scientific, 1996, pp. 43-62.

[14] M.R. Mehta, C. Dasgupta and G.R. Ullal, "A neural network model for kindling of focal epilepsy", in Neural modeling of brain and cognitive disorders, J.A. Reggia, E. Ruppin and R. Sloan Berndt (Eds.) Singapore: World Scientific, 1996, pp. 347-366.

[15] M. Raeiati Banadkooki, C. Mielke, K. H. Wolf, R. Haus and M. Marschollek, "Automatic detection of depression by using a neural network", Stud. Health Technol. Inform., 251, pp. 3-6, 2018.

[16] P. Lanillos, D. Oliva, A. Philippsen, Y. Yamashita, Y. Nagai and G. Cheng "A review on neural network models of schizophrenia and autism spectrum disorder", Neural Networks, 122, pp. 338-363, 2020. https://www.sciencedirect.com/science/article/pii/S0893608019303363T

[17] A.D. Redish and J. A. Gordon (Eds.), Computational psychiatry: new perspectives on mental illness. Cambridge: Mass.: The MIT Press, 2016.

[18] M. Sribala, "An approach of artificial neural networks for prediction of generalized anxiety disorder", IJRCAR, 3(3), pp.118-124, 2015.

[19] T.J. Meyer, M.L. Miller, R.L. Metzger and T.D. Borkovec, "Development and validation of the Penn State Worry Questionnaire," Behaviour Research & Therapy, 28, pp. 487-495, 1990.

[20] B. Sandín, P. Chorot, R.M. Valiente and L. Lostao, "Validación española del cuestionario de preocupación PSWQ: estructura factorial y propiedades psicométricas," Revista de Psicopatología y Psicología Clínica, vol. 14(2), pp. 107-122, 2009.

[21] S.E. Fahlman and C. Lebiere, "The cascade-correlation learning architecture," in Advances in neural information processing systems 2, D.S. Touretzky, Ed., Los Altos, Calif.: Morgan Kaufmann, 1990, pp.524–532.

[22] L. Pretzchel, "Early stopping-but when?, in Neural networks: tricks of the trade, G. Montavon, G. Orr and K.-R. Müller (Eds.) Berlin, Springer, 2012, pp.53–67.

[23] MEMBRAIN - NN.DE. https://membrain-nn.de/main_en.htm.

[24] A. Popko, M. Jakubowskl and R. Wawer, "MemBrain neural network for visual pattern recognition," ASTRJ, 7(18), June 2013, pp. 54-59. https:// doi-org/:10.5604/20804075.10512571990.

[25] A. Popko, "Optical horizon recognition using artificial neural networks," ICCSEAS, November 2017, pp. 33-38, 2017.

[26] S. Haykin, Neural networks: A comprehensive foundation, 2nd ed. Upper Saddle River, N.J.: Prentice-Hall, 1999.

[27] T.R. Shultz, Computational developmental psychology. Cambridge, Mass.: The MIT Press, 2003.

[28] E.A. Selby, M.D. Anestis and T.E. Joiner, Jr., "Understanding the relationship between emotional and behavioral dysregulation: emotional cascades," Behaviour Research & Therapy, 46, pp. 593-611, 2008. https://doi.org/: 10.1016/j.brat.2008.02.002.

[29] C. Papageorgiou and A. Wells, Eds., Depressive rumination: Nature, theory and treatment. Chichester: John Wiley&Sons, 2004.

[30] E.A. Selby and T.E. Joiner, Jr., "Emotional cascades as prospective predictors of dysregulated behaviors in borderline personality disorder," Personality Disorders: Theory, Research and Treatment, 4(2), pp. 168-174, 2013. https://doi.org/10.1037/a0029933.

Performance Comparison of CRUD Methods using NET Object Relational Mappers: A Case Study

Doina Zmaranda[1], Lucian-Laurentiu Pop-Fele[2], Cornelia Győrödi[3], Robert Győrödi[4], George Pecherle[5]
Department of Computer Science and Information Technology, University of Oradea, Oradea, Romania[1, 3, 4, 5]
Faculty of Electrical Engineering and Information Technology, University of Oradea, Oradea, Romania[2]

***Abstract*—Most applications available nowadays are using an Object Relational Mapper (ORM) to access and save data. The additional layer that is being wrapped over the database induces a performance impact in detrimental of raw SQL queries; on the other side, the advantages of using ORMs by focusing on domain level through application development represent a premise for easier development and simpler code maintenance. In this context, this paper makes a performance comparison between three of the most used ORM technologies from the .NET family: Entity Framework Core 2.2, nHibernate 5.2.3 and Dapper 1.50.5. The main objective of the paper is to make a comparative analysis of the impact that a specific ORM has on application performance when realizing database requests. In order to perform the analysis, a specific testing architecture was designed to ensure the consistency of tests. Performance evaluation for time responses and memory usage for each technology was done using the same CRUD (Create Read Update Delete) operations on the database. The results obtained proved that the decision to use one of another is dependent of the most used type of operation. A comprehensive discussion based on results analysis is done in order to support a decision for choosing a specific ORM by the software engineers in the process of software design and development.**

Keywords—ORM (Object Relational Mapper); domain-level development; performance evaluation; CRUD (Create Read Update Delete) operations

I. Introduction

ORM (Object Relational Mapper) is a pattern for accessing a relational database from an object-oriented language, with several implementation for almost every language. Basic features of an ORM include support for at least one specific persistence engine and CRUD operations. Some ORM features also include custom-SQL extensions for query building. Consequently, an ORM is a library that uses the object-oriented paradigm in a specific language to write a query that will return a set of data mapped into an object type that is needed [1].

Generally, the use of ORMs will have a negative impact on the application's execution time because it is being wrapped over the relational database and, in comparison to raw SQL queries, which can be stored as procedures or functions in a database, it provides slower return time values for all types of requests. However, ORM libraries could be preferred by developers in detrimental of raw SQL queries because of the easiness of writing data accessing code, faster debugging, ORMs being more readable than raw SQL [2], thus resulting better quality software. Nowadays, there are many ORM libraries free of charge or with a paid license which are offering a great set of functionalities ready to be used out of the box and with constant updates. The main advantage of using an ORM is represented by the fact that development is focused on the domain (model) level that describes at higher level of abstraction how program data is stored and retrieved from the database, leading to easier development and code maintenance. On the other side, by introducing an additional layer, performance issues arise. Depending on the ORM, this performance downsides are introduced either by internal way of entity-model design approach or, by using reflection [3].

From several ORMs that exists nowadays, this paper focuses on the three of the most used ORMs for .NET applications development: Entity Framework Core [4], nHibernate [5] and Dapper: EF Core 2.2.3 with EF Core Proxies 2.2.3 and EF Core SqlServer 2.2.2 libraries alongside with Dapper 1.50.5 library and using Dapper.Bulk 1.4.2 for bulk operations and nHibernate 5.2.3, configured with FluentNHibernate 2.1.2. A complex analysis and comparison between these ORMs impact on application's data interrogation methods performance is presented in the paper, by analyzing multiple CRUD calls with different levels of complexity. The main objective is to provide an overall experimental study that helps developers and architects when considering the trade-off between benefits of ORMs and their performance drawbacks when developing an application.

Execution time and memory footprint are considered the metric to realize the comparisons; a specific testing architecture was developed for running the tests and comparing ORM's performance results depending on different CRUD operation. This architecture implies the development of an application that targets a custom-made database and uses a specific benchmarking library, DotNetBenchmark [6], together with ORM's specific-developed repository class, to test the execution time and memory usage of multiple CRUD operations, as well as testing on multiple runtimes.

The paper is organized as following: in the first chapter, a short introduction emphasizing the motivation of the paper is presented, followed by chapter two that reviews related work. The method and testing architecture are illustrated in chapter 3 and the obtained experimental results are presented in chapter 4. An overall analysis and discussion about the results is described in chapter 5 and finally some conclusions are drawn.

II. Related Work

Several comparisons were done in the literature between different .NET ORM technologies, but they are generally targeting only two at the time. Solutions are generally analyzed in terms of performance, as in [7], but also in terms of their impact on application development [8].

Translation overhead for persistence operation is analyzed in [9] from the perspective of the additional layer introduced by the ORMs, by making a comparative analysis from the software development point of view of the two most used ORM tools in .NET programming environment: Entity Framework and nHibernate. Other studies analyze the benefits of using an ORM versus the drawback performance induced by ORM [10], by giving an insight look to the generated code. A study of the performance of Entity Framework and nHibernate for different types of databases (MS SQL Server and PostgreSQL) and using different query languages (lambda expressions and LINQ for Entity Framework and HQL and Critera API for NHibernate) in comparison with using SqlClient queries is presented also in [11]. Another arising issue is related to energy efficiency of ORM approaches, as it is described by the authors in [12], a study that experimentally evaluates energy efficiency of three different approaches to programmatically access SQL databases from PHP applications.

When coming to the recent versions of the .NET ORMs practical performance issues (caching, lazy loading, future queries) when building robust and scalable data access layer using NHibernate's are described in [13]; also, in [14] an approach for detection of ORM performance anti-patterns in the source code regarding database access details is presented. From the performance point of view, in [15] a fetch performance comparison by conducting experiments on common test data set of selected data access libraries: ADO.NET, Dapper and Entity Framework Core with tracking and no-tracking change is investigated. A common conclusion that results is that generally, using ORMs for application development introduce several benefits when compared to a plain SQL approach. On the other side, these techniques have well known disadvantages; but, as outlined in [8], the latest versions of skilfully developed ORMs is likely to generate well-tuned code that minimizes the performance impact on modern applications. The simpler ORMS, as Dapper, tend to work faster but exhibits fewer functionalities than the most complex ones, like EF Core or NHibernate. However, since performance issues could depend on the type and complexity of operation and on the volume of entries, none of these studies presents a comparison between all three and analyze multiple CRUD calls with different levels of complexity.

III. Method and Testing Architecture

The method used for testing performs experimental tests for all three different ORMs: Entity Framework Core, nHibernate and Dapper. For each ORM, different type of queries (Insert, Get, Update and Delete) with different degree of complexity were run on the same database. Execution time and memory consumption were monitored over different number of entries implied in the operation. The testing architecture, used to realize the comparisons in the present study, is presented in Fig. 1. The testing architecture implies a custom developed targeted database used for benchmarking testing.

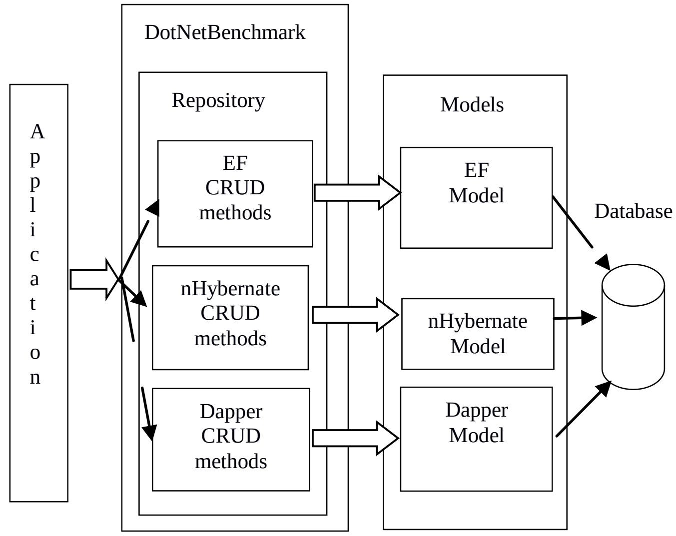


Fig. 1. Testing Architecture.

In order to separate the main concepts for each ORM technology, a project was developed as a Console Application in Visual Studio 2017 Professional. Furthermore, DotNetBenchmark specialized library was used by the application for benchmarking on the database by using different CRUD operations on multiple runs [6]. An Intel Core i7-6700HQ CPU 2.6GHz (Skylake) with 8 logical and 4 physical cores, 16 GB RAM and 256 GB SSD was used as underlying hardware support. It runs Windows 10 Pro operating system. For this purpose, a repository class was created inside the application for defining the CRUD methods for each ORM. Finally, all CRUD operations were tested through this unique application in a consistent way.

A. The Database

MSSQL Server 2018 and SQL Server Management Studio were used for direct access and table visualization in the development process. The database schema can be observed in Fig. 2; it was created so that it includes all types of table relationships (one-to-one, one-to-many and many-to-many), in order to be able to test and observe the performance impact of those relation types on different operations.

B. Application Project Structure

An application project *UniversityDBenchmark* was created for analysing the three different ORM technologies (Dapper, EntityFramework Core 2.2 and nHibernate); application structure was designed specifically to separate the main concepts for each used technology. The project structure is presented Fig. 3a and consists of the following modules:

- Benchmark–represents container with a class developed to configure the methods used for benchmarking
- Context–a container with a class used to configure the Entity ramework context class (UniversityContext.cs)

- Helpers–container for a class that is returning the database connection string from appsettings.json file
- Logger–container for class that provides logic to log the SQLs generated by the entity framework methods
- Repository–a container with repository classes that contains CRUD methods for accessing the database for each ORM
- Program.cs–represents the main gate to the Console application, used to call to run benchmark analysis methods

C. Benchmark Analysis

The benchmarking was realized using the specialized library DotNetBenchmark which is offering means of testing the execution time of multiple methods as well as testing different runtimes, as described in [6]. The BenchmarkAnalysis class provides also access to each repository CRUD methods that will be dynamically called; each method was marked with the [Benchmark] attribute, by this approach telling the library which methods to include in the current benchmark session (Fig. 3b). The RankColumn defines a column in the results table, after the benchmarking process has been finished, that contains the execution time ranks in ascending order of the declared methods.

MinIterationCount/MaxIterationCount attributes are forcing the benchmarking library to execute between 10 and 20 times each method. MemoryDiagnoser is an attribute by which memory surveillance during the actual method calls is enabled, so that in the results table, the memory used to execute these methods will be shown. MinColumn and MaxColumn have the purpose of displaying the observed minimum and maximum analysis values for each tested method that will be used to compute the average final benchmark analysis value. An example about how to declare the repository methods so that it will be included into benchmarking analysis is the following (for InsertStudents method):

```
[Benchmark]
   public void InsertStudentsWithNHibernate() =>
_NHibernateRepository.InsertStudents(iterationNumber);

[Benchmark]
public void InsertStudentsWithEF() =>
   _EFRepo.InsertStudents(iterationNumber);
[Benchmark]
 public void InsertStudentsWithDapper() =>
   _DapperRepo.InsertStudents(iterationNumber);
```

For all three ORMs used for testing, the repository class implements all specific CRUD methods targeting the same tables from the database. Thus, a series of CRUD calls to the database using each targeted technology (Entity Framework Core, nHibernate and Dapper) were created, with two types of method calls: one for simple scenarios (in which 3 tables in one-to-one relationship, having 2 one-to-one relationships are targeted) and one for a more complex scenario (in which 4 tables which are in one-to-one and one-to-many relationship were targeted - 3 tables linked by 2 one-to-one relationship and 2 tables linked by a one-to-many relationship). In order to obtain the results, all these methods will be called by a defined number of times: 500, 1000, 2000, 5000 and 10000. The targeted operations are:

- INSERT–with the corresponding methods insert students and insert teachers with affiliated courses
- GET–with the corresponding methods getting a number of students and teachers and getting all students participating to a teacher's courses. This method, was run on a high number of entities
- UPDATE–with the corresponding methods update student's address and update teacher's address and courses description; before the actual update a get to take a specific number of students/teachers is run, these calls being also monitored and decreased from the overall update execution time
- DELETE–with the corresponding methods delete students and delete teachers and corresponding courses; in this case a get method to take a specific number of entries which will be targeted for deletion was run prior to deletion and its execution time was decreased from the overall deletion execution time.

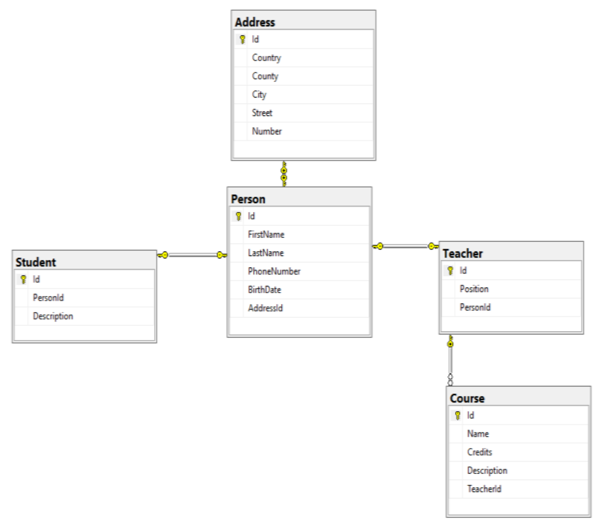


Fig. 2. University Database Schema.

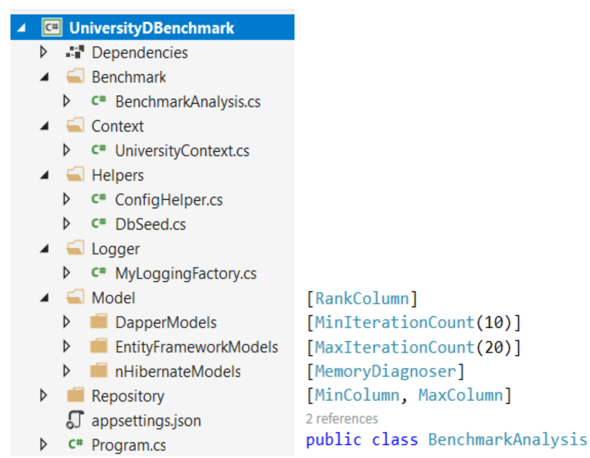


Fig. 3. (a) Project Structure. (b) Benchmark Attribute.

IV. Experimental Results

Performance comparison was using DotNetBenchmark benchmarking library for execution time and memory allocation for each called method. Each framework will store in memory the results after running the generated SQL statements. The logic that translates the methods into equivalent SQL will also cost memory. EF Core for example is using reflection and thus, the first call will be slower and more time consuming than the rest of them because of the caching used. Dapper is not using reflection but is having logic for mapping results from the SQL that was generated and ran and the expected entities. nHibernate is also using first level of cache for optimization, but the mappings behind the scenes will put a mark on the memory consumption.

A. Insert

Two scenarios were used for testing INSERT operation with the targeted technologies. First scenario represented by insert students method implies inserting entities that have two one-to-one relationship (Student-Person-Address) that means the logic will add first an address, person and finally a student. The second scenario, represented by insert teachers with affiliated courses method, implies inserting entities which have besides the one-to-one relationship also one-to-many (Teacher-Person-Address, Teacher->Courses), that means the logic will add first an address, person, teacher and then a list of courses for each teacher entity.

The insert method used for the first insertion scenario (*insert students*) is the same for all three technologies; the one-to-one relationship between Student-Person-Address implies that 3 inserts for each student entity insert process will be made:

foreach Student
create Address;
create Person;
add Address to Person
create Student;
add Person to Student
save Student;

However, some particularities were considered for each ORM:

- Entity Framework Core: since behind the scenes it uses transactions, there is no need to include the simple save statement into a new transaction;
- nHibernate: in order to make calls to the database using this technology it is necessary to manually open a session
- and a transaction and all operations to be made inside this opened transaction. After each transaction has ended, it is recommended to clear (flush) the session;
- Dapper: to benefit of using a library that is adding bulk data, the approach of separately adding entities was chosen.

The obtained results after the executing the benchmark analysis are represented in Fig. 4 and Table I. Is obviously that the best timing results for insert students scenario are obtained by nHibernate, mainly because of its simple underlying logic and also because of the simplicity of the next SQL statement which is called after an INSERT, that for nHibernate has a simple form: SELECT scope_Identity().

Dapper is also very close to nHibernate's results, being a bit slower because under the hoods, Dapper is running a lightly different SQL in the form of: SELECT CAST (scope_identity()), and that particular cast operation will mark its effect upon the total result. EF Core is on the last position because, after the INSERT, it is running a much more complex SQL query: SELECT where @@rowcount = 1 and [id] = scope_identity(), which is time-costly compared to the other two.

For the second insertion scenario (represented by insert teachers with affiliated courses) for each teacher inserted, a new address, a new person and a list of courses will be added (one-to-one/one-to-many relationships). Consequently, when adding a new teacher, a new address and a new person entity are required. The insert method used is the following:

foreach teacher create Address;
create Person;
add Address;
create Teacher;
add Person;
create Courses;
add Courses to Teacher;
save teacher;

The obtained results are presented in Fig. 6 and Table I.

When adding more complex entities, nHibernate has still be best values than EF Core and Dapper, just the same as for the last scenario. Thus, from the time results that overall, nHibernate would be the best option for realising *Insert* operations followed closer by Dapper; if memory usage is considered, the best is still nHibernate followed by EF Core. Memory allocation for each technology is presented in Fig. 5 and Table II. An explanation for EF Core time results is that behind the scene, is adding the Teacher entity and then all courses are being saved into a temporary table that is merged with the Course table using a select on Course table joined with the temporary table. In comparison, Dapper and nHibernate are realizing simple inserts with the given values.

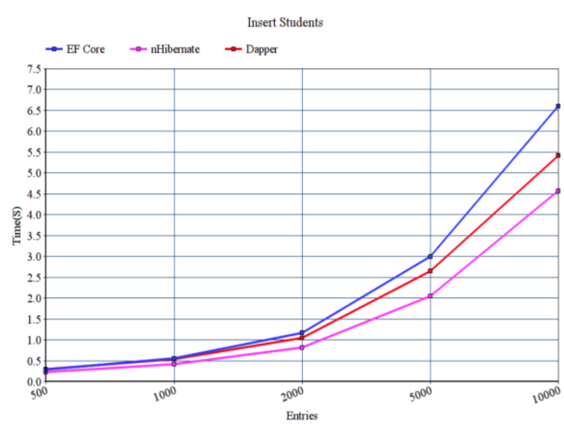


Fig. 4. Insert Students (One-to-One).

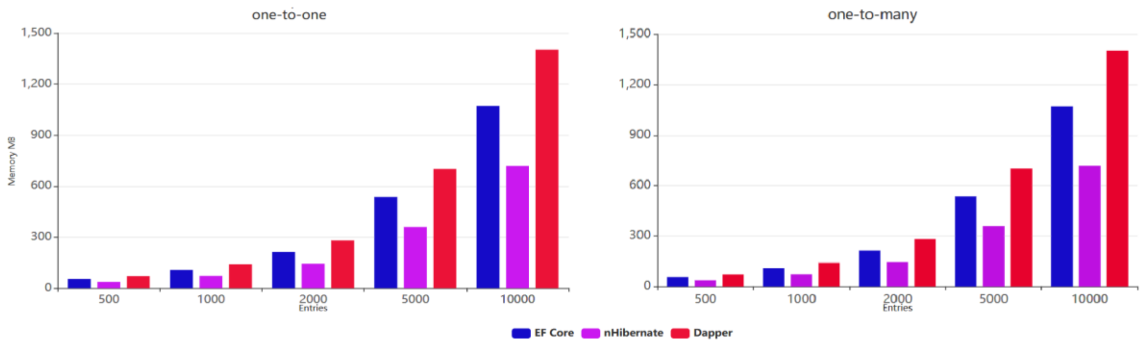


Fig. 5. Memory used – Insert.

TABLE. I. INSERT METHOD EXECUTION TIME

Entries /Time(s)	**Insert methods Time(s)**					
	Tab Entity Framework Core		***nHibernate***		***Dapper***	
	Students	***Teachers***	***Students***	***Teachers***	***Students***	***Teachers***
500	0.28	0.82	0.22	0.56	0.29	0.67
1000	0.55	1.44	0.41	1.24	0.53	1.51
2000	1.16	3.21	0.81	2.29	1.04	2.56
5000	2.99	7.28	2.04	5.92	2.64	7.98
10000	6.6	15.01	4.75	11.76	5.41	13.65

TABLE. II. INSERT METHOD MEMORY USAGE

Entries /MB	**Insert methods memory usage (MB)**					
	Tab Entity Framework Core		***nHibernate***		***Dapper***	
	Students	***Teachers***	***Students***	***Teachers***	***Students***	***Teachers***
500	21.93	53.45	13.1	29.16	30.83	66.68
1000	43.83	106.74	26.2	57.9	61.65	133.34
2000	86.8	12.85	52.43	115.88	123.28	266.67
5000	217.45	535.5	130.66	291.04	261.16	666.65
10000	435.57	1071	261.16	582.05	616.35	1333.29

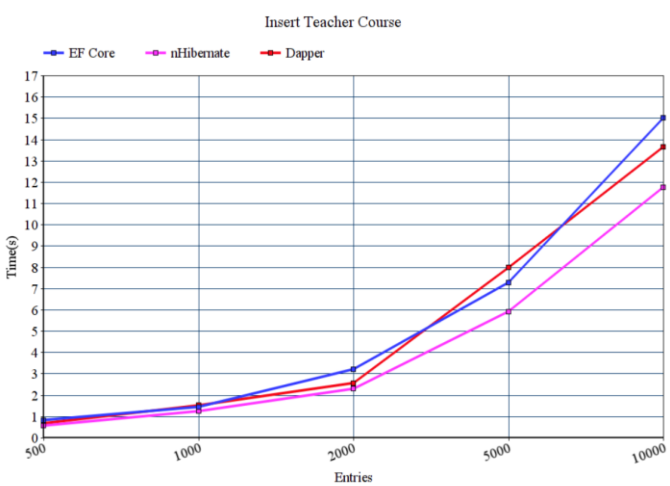


Fig. 6. Insert Teachers with Corresponding Courses (One-to-many).

B. Update

The same testing approach from Insert was applied also to Update methods, by using two scenarios: updating students addresses and updating teacher's address and corresponding courses in order to test the update process of entries being in a one-to-one respectively one-to-many relationship. Each method will update a specific number of students and teachers that will be updated: 500, 1000, 2000, 5000, 10000. The database on which update statements were called has a constant 10.000 entries in Student, Person and Address table for the first scenario and a constant of 10.000 entries in Teacher, Person and Address and 40.000 in Course table for the second one. During benchmarking process, the Get methods will also be considered so that the exact time for the update statement alone is computed. For the first scenario of updating student's addresses, the same update statement is used for all three technologies:

get Students (with Person and Address)
foreach Student
    change Address; update Student

All three technologies have the same logic flow, with minor differences:

- EF Core: uses the Include option to retrieve the related one-to-one entities; afterwards, a simple build-in Update method is called upon the retrieval of Student entities which will detect all changes then update them accordingly. Even there are multiple Update calls, all updates effectively take place only when context's SaveChanges() method is hit;
- NHibernate: just like EF Core, firstly it retrieves a given number of Students alongside with all related entities by using the call QueryOver; then each student's address will be changed, and the build-in Update call is triggered. Just as for EF Core, there are multiple Update calls, but only when committing the opened transaction, the statements will be triggered;
- Dapper: in particular for Dapper, all the auto-generated SQL calls that EF Core is generating behind the scenes when is getting the Student entity were re-created in order to provide a much more accurate analysis between these technologies. After getting all students and related entities, the addresses were modified then the Update method was called.

The obtained results after the execution time benchmark analysis of updating student's addresses are presented in Fig. 8 and Table III. The best timing results for *update students* scenario are obtained by EF Core that uses by default eager loading when nHibernate is using lazy loading by default: when realising a student entity update, firstly it is needed to return from database all data that did not have yet been loaded. EF Core has better timing results also in comparison with Dapper mainly because of the underlying logic from EF Core Update method, because from the code perspective, EF Core Update Students and Dapper Update Students are logically the same.

The second update scenario is approaching to update entities via both one-to-one and one-to-many relationships. In this case, Teacher's table is selected, which is linked with a one-to-one relationship with Person that is linked with Address also with a one-to-one relationship just as for Student table. Besides this link, it has a one-to-many link to Course table. During this scenario tests, the Address and a Course from each Teacher considered will be changed. A get statement needs to be run first in order to obtain all Teachers that need update and simulate through this multiple update calls.

get Teachers;
foreach Teacher get Courses;
foreach Teacher;
update Address;
update Course.Description; update Teacher;

As it results from Fig. 9 and Table III, all three technologies are having higher time results for updating teacher's courses in comparison with updating the Student table, but nHibernate has the biggest time result from all three. This can be explained also by the fact that nHibernate is using by default lazy loading and so, even if all teachers were returned with an initial *Get* call (simple Select from database) a call the database every time other inner-entities from Teacher object are accessed is needed.

Therefore, the overall timing for the nHibernate update call is increasing. Entity Framework and Dapper are using eager loading and therefore they do not need to access the database each time an object which was not previously loaded is changed. Nevertheless, EF Core, in this case, is slightly overcome by Dapper but the difference is relatively small.

The update methods memory allocation for each technology used is presented in Fig. 7 and Table IV. For the one-to-one approach, the the consequences of using lazy loading by default for nHibernate can be seen, that implies extra memory usage when realising the Update statement because of the need to return and save the entities in the same Update call before saving the changes.

When also one-to many relationships are involved, EF Core and Dapper are having higher memory usage when updating entities than nHibernate because of the used underlying logic. Consequently, EF Core is generally having the best results from both time and memory perspectives followed closely by Dapper, except the situation of memory usage, where slightly higher memory consumption can be seen when updating entities with one-to-many relationship to EF Core.

TABLE. III. UPDATE METHOD EXECUTION TIMES

Entries /Time(s)	Update methods Time(s)					
	Tab Entity Framework Core		*nHibernate*		*Dapper*	
	Students	*Teachers*	*Students*	*Teachers*	*Students*	*Teachers*
500	0.081	0.703	0.176	2.61	0.175	0.45
1000	0.191	1.625	0.361	5.78	0.339	0.88
2000	0.38	3.23	0.726	11.56	0.654	1.75
5000	1.03	7.36	2.03	25.27	1.66	2.48
10000	1.85	14.53	3.96	56.31	3.51	5.12

TABLE. IV. UPDATE METHOD USED MEMORY

Entries /MB	Update methods used memory (MB)					
	Tab Entity Framework Core		*nHibernate*		*Dapper*	
	Students	*Teachers*	*Students*	*Teachers*	*Students*	*Teachers*
500	7.6	52	19.2	35	8.84	60.24
1000	15.2	104.72	38.4	69.61	17.68	120.46
2000	30.4	211.57	76.8	138.77	35.37	240.97
5000	76	537.62	192	494.56	88.67	602.28
10000	159	1075.24	374	989.12	177.08	1125.93

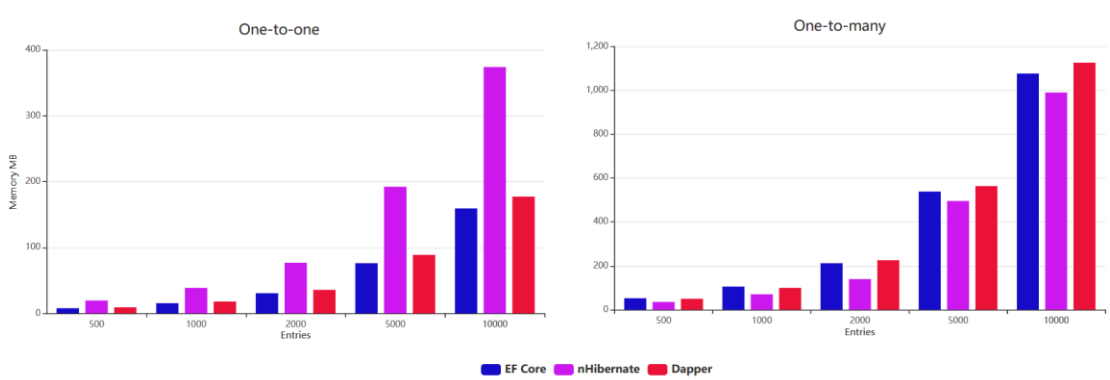


Fig. 7. Memory used – Update.

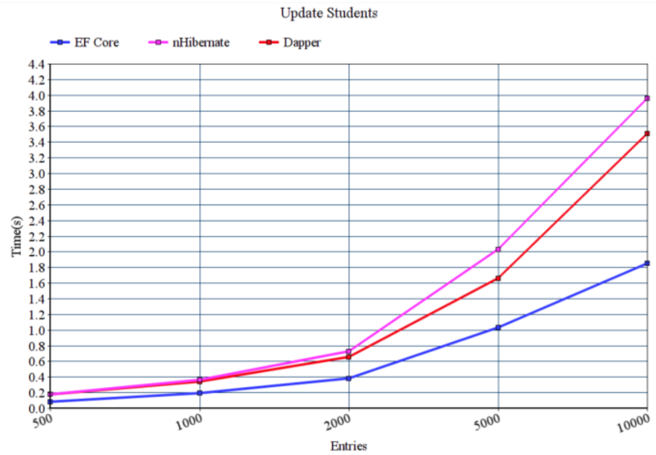


Fig. 8. Update Students (One-to-One).

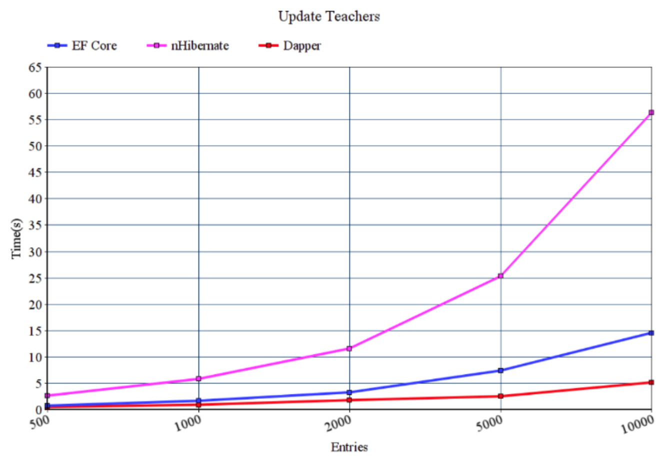


Fig. 9. Update Teachers with Corresponding Courses (One-to-Many).

For nHibernate in particular, as it is using lazy loading by default when realising an Update, if the targeted entity is not yet loaded, it will make a Select statement first, the will update the returned entity's values with the new ones, then will realise the UPDATE on the database. All these operations are very costly from the execution time point of view.

When looking at memory usage, a considerable difference when updating entities with one-to-one relationship versus one-to-many can be observed. This could be explained by the underlying logic which translates the .NET code into SQL statements because, when checking with SQL Profiler, in all cases, simple SQL *Update* statements are being made.

C. Delete

Delete operation follows the same approach by taking into consideration two scenarios: delete students and delete teachers and corresponding courses. Delete students implies deleting entries from student, person and address tables involving only one-to-one relationships; delete teachers implies deleting entries from teacher, person and address tables including corresponding courses (both on-to-one and one-to-many relationships). As in the previous methods that were analysed, a Select statement to get top 500/1000/2000/5000/10000 entries from Student and Teacher tables with all related entries will be previously run. The select queries will be then measured and extracted from the total results, after a delete process, in order to obtain only delete operation values. For the first scenario of deleting all students, the same delete statement is used for all three targeted technologies:

get Students and related entities;
for each Student
    delete Address; delete Person; delete Students;
end;

Just as for *Insert* and *Update* cases, also *Delete* operation has some particularities depending each targeted technology:

- EF Core: because of the foreign key AddressId present in Person table and PersonId present in Student table, the Address table is seen as the parent, so deleting the Address will automatically cascade delete also to the corresponding Person and Student
- NHibernate: because of the mapping classes, entities for cascade deletion can be directly marked and consequently, by simply deleting the teacher entity, all other related entities will be deleted; this behaviour is different when compared to EF Core and Dapper where in order to trigger the cascade deletion it is needed to delete the parent entity;
- Dapper: the simple Delete method from SimpleCRUD Dapper library was used to remove the parent Address entity and the BulkDelete method from Dapper-PLUS library was used to remove all related Teacher Course entities just as for EF Core case.

Before making the deletion statements, a *Select* query will be run to return a top 500/1000/2000/5000/10000 teachers from the database whose execution time was measured so that only the delete statement time for each used technology could be computed. From Fig. 11 and Table V it is obvious that EF Core is having the best timing results, close to Dapper, nHibernate having the worst time results in this case. NHibernate in this case is affected by the lazy loading default setting by having to load all yet unloaded entities before deleting them. Dapper and EF Core are having very close timing results, as the simplest delete option from each of them has been used, the difference between them could be explained by the logic behind the scenes which is transforming the code to SQL statements.

The second scenario used for deletion is running the Delete queries with the purpose of removing a given number of teachers with their corresponding courses. This will require a deletion of two one-to-one relationship (Teacher-Person-Address) and one one-to-many relationship (Teacher-Courses):

get Teachers and related entities;
foreach Teacher
    delete Address; *delete Person;*
    delete Teacher; *delete Courses (if any);*
end;

In this scenario, Dapper and EF Core have close results for delete statements and much higher than nHibernate for the case of deleting one-to-many entities, where nHibernate is having the best results in comparison with the case when is deleting only one-to-one relationship (Fig. 12 and Table V). This could be explained mainly because of the underlying logic for Dapper and EF Core that translates the Remove (EF Core) / BulkDelete (Dapper) methods into explicit delete SQL statement that could have an impact on execution time.

Dapper and EF Core are having better deleting timing results than nHibernate until reaching the level of 1000 deleted entries. By analysing the resulted SQL statements after a deletion command, the SQLs are very much the same for all three technologies, with the exception for EF Core that is running also a Select @@Rowcount to check return the number of effected entries. The delete methods memory allocation for each technology used is presented in Fig.10 and Table VI.

According to all these results, nHibernate has the highest memory usage when deleting entities in both one-to-one and one-to-many relationship. An explanation for this massive memory usage in comparison with EF Core and Dapper could be the fact that nHibernate is using lazy loading by default.

Consequently, instead of the real objects, it has some proxys that will be replaced with real object upon accessing their values.

Both EF Core and Dapper have good memory usage results, Dapper being more efficient than EF Core when deleting entities in one-to-many relationship, maybe because, behind the scenes EF Core and Dapper are having a better logic implemented than nHibernate and possible because EF Core and Dapper are using eager loading when returning entities without the need of the proxy objects that nHibernate is using to replace the objects not yet returned from the database.

D. Get

Get queries have been already run in the scenarios presented before to obtain a specific number of entities on which update or delete queries where run afterwards. Two scenarios were considered also here: getting a number of students and getting all students participating to a teacher's courses to return entities with one-to-one and one-to-many relationships respectively. These methods were run on a higher number of entities. Consequently, first the Get calls were made upon a database with 10.000 entries on Student table, then cleared the database and its cache and re-entered 10.000 Teacher entities alongside with 40.000 Courses (each teacher will have 4 courses). The is the following:

get Students (including Person and Address);

TABLE. V. DELETE METHOD EXECUTION TIMES

Entries /Time(s)	**Delete methods Time(s)**					
	Tab Entity Framework Core		***nHibernate***		***Dapper***	
	Students	***Teachers***	***Students***	***Teachers***	***Students***	***Teachers***
500	0.15	2.61	0.38	3.59	0.19	2.84
1000	0.28	8.43	0.72	6.36	0.31	9.34
2000	0.64	20.07	1.54	13.13	0.68	27.05
5000	1.49	39.12	2.82	26.72	1.74	53.75
10000	2.81	75.16	5.78	53.44	3.56	108.43

TABLE. VI. DELETE METHOD USED MEMORY

Entries /MB	**Delete methods used memory (MB)**					
	Tab Entity Framework Core		***nHibernate***		***Dapper***	
	Students	***Teachers***	***Students***	***Teachers***	***Students***	***Teachers***
500	12.78	4.21	27.98	87.31	3.98	8.53
1000	26.02	8.56	55.98	144	7.77	42.96
2000	52.26	17.14	112.98	289	15.56	85.93
5000	131.26	34.8	279.9	720	38.85	214.8
10000	262.52	69.6	559.9	1440	77.7	429.6

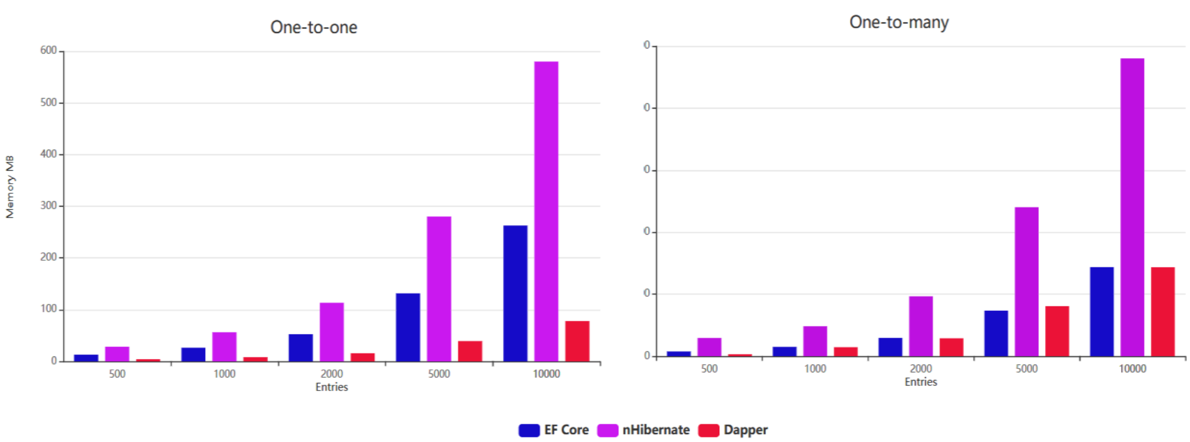


Fig. 10. Memory used – Delete.

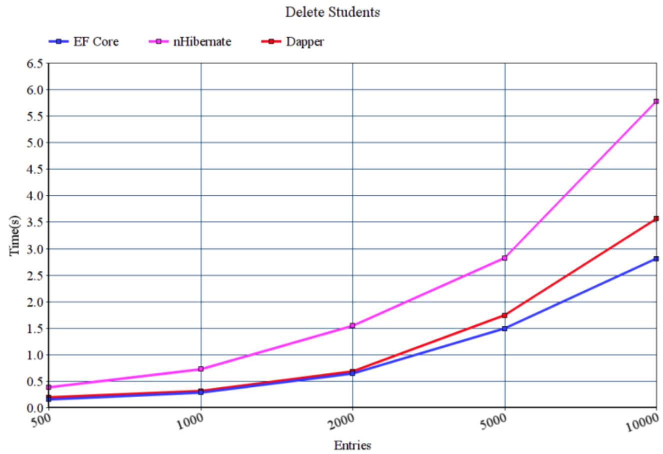


Fig. 11. Delete Students (One-to-One).

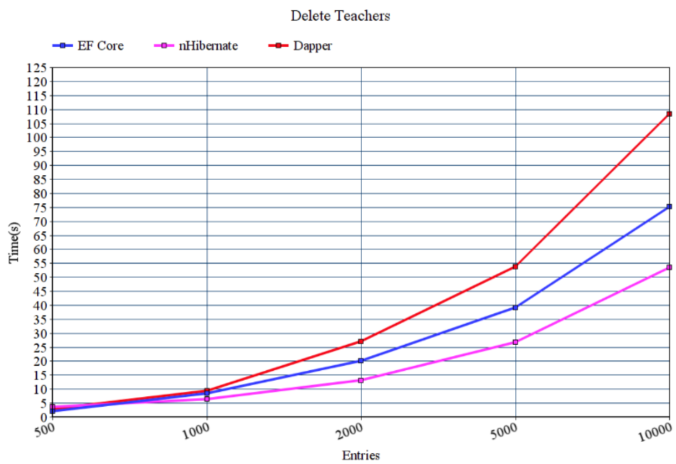


Fig. 12. Delete Teachers with Corresponding Courses (One-to-Many).

The Select (Get) operation has some particularities depending each targeted technology, that implies few differences as follows:

- EF Core: for data retrieval, the Include and ThenInclude functions were used, which will return multiple levels of related data also using the method Take() in order to select a given top entities from database. Eager loading is used (by default) when returning data, as lazy loading needs to be specifically turned on as from EF Core 2.1;
- Nhibernate: QueryOver<EntityType> generic method is used to retrieve all data regarding the given entity as parameter and also the Take() method to select a top from database. Lazy loading is used as it is the default behaviour for nHibernate.
- Dapper: the same SQL codes which EF Core is generating when running a Get Students was replicated used by injecting the SQL with Query<*EntityType*> method, to test the speed of Dapper when running and retrieving the result entities. Also, eager loading is used here, as Dapper being a direct-SQL library.

The obtained results are presented in Fig. 13 and Table VII. Dapper has the best timing results when returning entities on one-to-one relationship because of the missing entity tracking logic which is present behind the scenes for EF Core and nHibernate.

The second scenario used for *Get* method implies the retrieval of teachers alongside with their corresponding courses. A simple pseudocode to describe this can be seen below:

get Teachers(including Person and Address);
    get Courses;

The obtained results after the execution time benchmark analysis are presented in Fig. 14 and Table VII. Dapper has registered very inefficient results, having the worst results from all three technologies when it comes to returning a higher number of entities. One possible cause for this could be the use of the generated EF Core SQL into Dapper database call, indeed, and not using a Dapper built-in method to retrieve one-to-many entities. The reason for that is that, after testing Dapper's abilities to manage the same SQL code EF Core is generating behind the scenes, it was considered preferable not to use predefined Dapper get method.

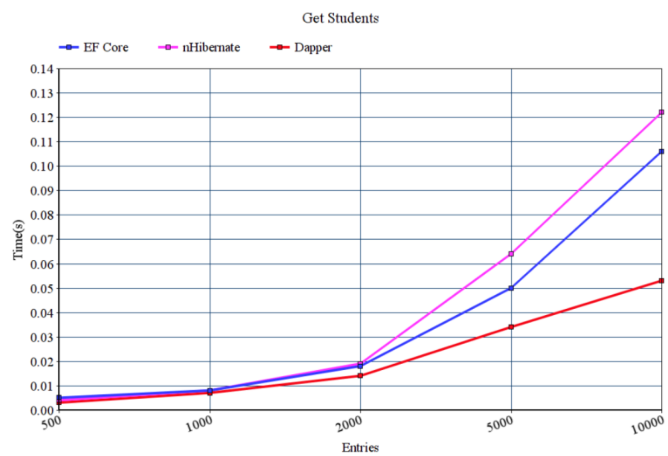


Fig. 13. Get Students (One-to-One).

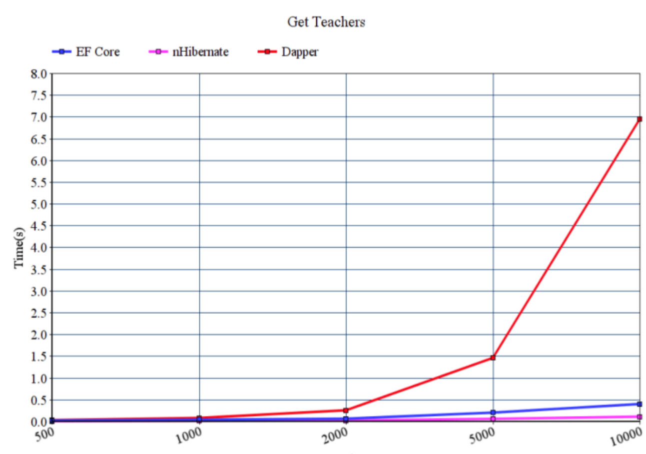


Fig. 14. Get Teachers with Corresponding Courses (One-to-Many).

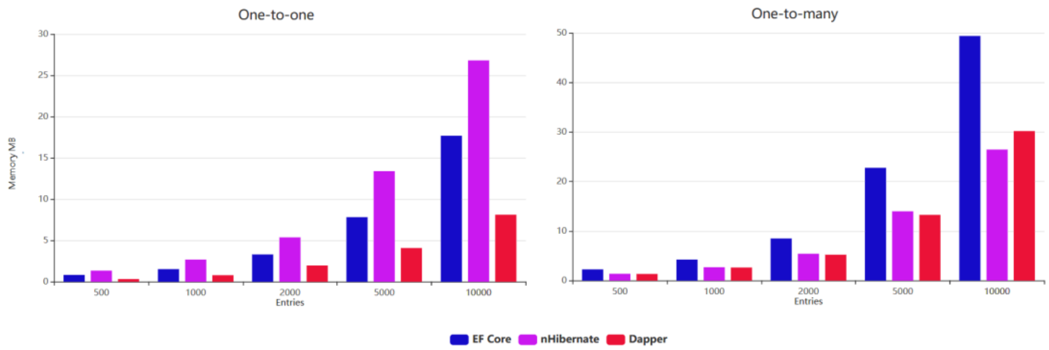


Fig. 15. Get Memory usage.

TABLE. VII. GET METHOD EXECUTION TIMES

Entries /Time(s)	Get methods Time(s)					
	Tab Entity Framework Core		*nHibernate*		*Dapper*	
	Students	*Teachers*	*Students*	*Teachers*	*Students*	*Teachers*
500	0.005	0.024	0.004	0.004	0.003	0.024
1000	0.008	0.037	0.008	0.011	0.007	0.042
2000	0.018	0.077	0.019	0.018	0.014	0.080
5000	0.050	0.201	0.064	0.058	0.034	0.185
10000	0.106	0.444	0.122	0.110	0.053	0.343

NHibernate is using by default lazy loading and thus having low timing results when returning one-to-many results, but this will be compensated by the calls that will have to be realised when accessing entities that have not yet been loaded. When returning one-to-one results, it is close to EF Core, both being overcome by Dapper, thing that could be easily explained by the simple fact that Dapper is in fact a micro-ORM and thus not having all the backside logic a fully-featured ORM has behind the scene. However, EF Core using eager loading has very close results to nHibernate in comparison to Dapper, but still 3 times slower than nHibernate.

By analysing the memory results in the one-to-one approach, it can be seen that nHibernate using proxy objects when returning entries using lazy loading (Fig. 15 and Table VIII).

TABLE. VIII. GET METHOD USED MEMORY

Entries /MB	Get methods used memory (MB)					
	Tab Entity Framework Core		*nHibernate*		*Dapper*	
	Students	*Teachers*	*Students*	*Teachers*	*Students*	*Teachers*
500	0.83	2.61	1.34	1.35	0.31	2.67
1000	1.51	4.98	2.68	2.69	0.78	5.32
2000	3.29	9.95	5.36	5.39	1.96	10.64
5000	7.81	26.43	13.4	13.97	4.07	26.8
10000	17.69	53.07	26.8	26.44	8.12	53.59

But, when selecting more complex entities (one-to many approach), EF Core and Dapper have higher memory usage for returning one-to-many entities, in comparison with the case of returning one-to-one entities, because of the impact of the logic used there to map all returned courses to each teacher while nHibernate is using proxy objects.

V. RESULTS AND DISCUSSION

For Insert operations, nHibernate has the best results, followed by EF Core and then by Dapper. For Update operations EF Core has the best results when considering one-to-one relationships; but when considering also one-to-many relationships, Dapper has the best results.

A similar situation was observed for the Delete operations, where, when considering one-to-one relationships, EF Core has best execution time results followed relatively closely by Dapper and then by nHibernate; but, when considering also one-to-many relationships, nHibernate has on average the best results in terms of execution time for high number of entries (over 1000) followed closely by EF Core and then by Dapper. This performance is obtained to the detriment of memory usage.

When realizing Get calls, nHibernate and EF Core execution times are very close; for one-to-one relationships, Dapper is having better results as number of entries increases, followed by EF Core and then by nHibernate; but, when also one-to-many relationships are involved, nHibernate has, on average, the best results in terms of execution time followed by Dapper, EF Core having the worst performance as the number of entries increases. Significant differences could be observed for Get operations from an overall perspective.

If using nHibernate's default settings, it will make use of lazy loading and thus showing great timing results and memory when one-to-many relationships are involved; but, for only one-to-one relationships its performance is the lowest one. EF Core with lazy loading being disabled exhibits good results when returning entities in one-to-one relationship but memory and time consuming when returning complex double one-to-many entities.

In a real-life project, all technologies could be used just with the statements on which are showing the best results, such as: nHibernate for Insert, nHibernate (using default lazy loading) or EF Core for Get, EF Core or Dapper for Update and Delete.

VI. Conclusions

After analyzing the results, it can be concluded that none of the three technologies targeted for benchmarking analysis is having the very best results both from the time point of view and memory usage. The decision to use one of another is dependent of the most used type of operation.

This paper's work could be further developed by testing all three technologies over an Azure stored database. This could bring another live scenario to test, when there is a need to measure time responses of queries that target a remote database. Nevertheless, this approach will definitely have its downsides, for example, the internet connection stability, bandwidth, database type chosen from Azure, location of the Azure Storage in accordance with the server location on which the application is running, all will have a major role in realizing measurements.

References

[1] K. Roebuck, Object-Relational Mapping: High-impact Strategies - What You Need to Know. Samford, Australia: Lightning Source Publisher, 633p., 2011.

[2] P. Van Zyl, D.G. Kourie, A. Boake, "Comparing the performance of object databases and ORM tools", Proceedings of the 2006 Annual research conference of the SAICSIT on IT research in developing countries, SAICSIT 2006, Somerset West, South Africa, DOI: 10.1145/1216262.1216263, pp. 1–11, 2006.

[3] A. Haug, J. Arlbjorn, F. Zachariassen, J. Schlichter, "Master data quality barriers: an empirical investigation". Industrial Management & Data Systems, Emerald Group Publishing Limited, DOI: 10.1108/02635571311303550, 113 vol. 2, pp. 234-249, 2013.

[4] J.P. Smith, Entity Framework Core in Action. New York: Manning Publications Company, 520p., 2018.

[5] G. Liljas, A. Zaytsev, J. Dentler J., NHibernate 4.x Cookbook - Second Edition. UK: Packt Publishing, 448p., 2017.

[6] Khan O.M.A., Benchmarking .NET Core 2.0 applications. In: C# 7 and .NET Core 2.0 High Performance: Build highly performant, multi-threaded, and concurrent applications using C# 7 and .NET Core 2.0. UK: Packt Publishing Ltd., 281p., 2018.

[7] Basheleishvili I., A. Bardavelidze, and K. Bardavelidze. "Study And Analysis Of The .Net Platform-Based Technologies For Working with the Databases, " Proceedings of the 33rd International Conference on Information Technologies (InfoTech-2019), Bulgaria, pp.1-8, 2019.

[8] V.J. Mehta, ORM Patterns and Domain-driven Design. In: Pro LINQ Object Relational Mapping with C#. New York: Apress Publisher, pp.17-43, 2008.

[9] S. Cvetković, D. Janković, "A Comparative Study of the Features and Performance of ORM Tools in a .NET Environment", in: A. Dearle, R.V.Zicari, Objects and Databases, ICOODB 2010. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, DOI: 10.1007/978-3-642-16092-9_14, vol. 6348, pp.147-158, 2010.

[10] A. Joshi, S. Kukreti, "Object Relational Mapping in Comparison to Traditional Data Access Techniques", International Journal of Scientific & Engineering Research Vol. 5, Issue 6, pp.540-543, 2014.

[11] A. Gruca, P. Podsiadło, "Performance Analysis of .NET Based Object–Relational Mapping Frameworks", 10th International Conference, BDAS Ustron, Poland, Springer International Publishing, DOI: 10.1007/978-3-319-06932-6, pp.40-49, 2014.

[12] G. Procaccianti, P. Lago, W. Diesveld, "Energy Efficiency of ORM Approaches: An Empirical Evaluation", Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement ESEM, Ciudad Real, Spain, ACM, DOI: 10.1145/2961111.2962586, pp.185–198, 2016.

[13] Suhas Chatekar, Learning NHibernate 4: Explore the full potential of NHibernate to build robust data access code. 402p. UK: Packt Publishing, 2015.

[14] T. H. Chen, W. Shang, Z. M. Jiang, A.E. Hassan, M. Nasser, P. Flora, "Detecting performance anti-patterns for applications developed using object-relational mapping.", Proceedings of the 36th International Conference on Software Engineering. ACM, pp.1001–1012, DOI:10.1145/2568225.25682592014, 2014.

[15] W. Wiphusitphunpol, T. Lertrusdachakul. "Fetch performance comparison of object relational mapper in. NET platform." In 2017 IEEE 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology ECTI-CON, DOI: 10.1109/ECTICon.2017.8096264, pp. 423-426, 2017.

Using the Convolution Neural Network Attempts to Match Japanese Women's Kimono and Obi

Kumiko Komizo[1], Noriaki Kuwahara[2]
Graduate School of Engineering and Science
Kyoto Institute of Technology
Kyoto, Japan

***Abstract*—Currently, the decline in kimono usage in Japan is serious. This has become an important problem for the kimono industry and kimono culture. The reason behind this lack of usage is that Japanese clothing has many strict rules attached to it. One of those difficult rules is that kimonos have status, and one must consider the proper kimono to wear depending on the place and type of event. At the same time, the obi (sash) also has status, and the status of the kimono and obi must match. The matching of the kimono and obi is called "obiawase" in Japanese, and it is not just a matter of the person wearing the kimono selecting a pair that she likes. Instead, the first place you wear a kimono determines its status, and the obi must match that status and kimono. In other words, the color, material, meaning behind the pattern must be matched with obi. Kimono patterns may evoke the seasons or a celebratory event. All this must be considered. The kimono was originally everyday wear, and people were taught these things in their households, but with today's increasingly nuclear families, that person who could teach these things isn't nearby, adding to the lack of use of kimonos. Because of this, there has been interest in using CNN (Convolution Neural Network) from the digital fashion industry. We are attempting to use machine learning to tackle the difficult task of matching an obi to a kimono, using the CNN machines drawing the most attention today.**

***Keywords*—Digital fashion; kimono; obi; convolution neural network**

I. Introduction

In Japan, the obi and kimono have changed little by little, influenced by social position, lifestyle, the tastes of the times, and fashion. In 2019, Japan began a new era called Reiwa. Since the end of World War II in 1945, it has been said that people have been wearing kimonos less and opting more for Western clothes [1]. Today in Japan, kimonos are mostly worn to celebratory occasions such as weddings and graduations. Before 1945, people in Japan wore kimonos every day, and they were free to match colors as they liked. However, with the outbreak of World War II, the wearing of slacks-style monpei pants and a government decreed national uniform began. The reasoning was that duing wartime, the kimono wasn't good for taking action, wasn't functional, and wasn't reasonable [1]. That's why the other type of clothing was accepted. Also, the fact that more households started to have sewing machines spurred a Western clothes boom, and more people began to wear Western clothes [2]. However, a certain number wore kimono to graduations and other celebratory events. According to the Kyoyuzen Komon Manufacturing Survey and Report, at the peak of manufacturing in 1971, Kyoyuzen's total production was 16,524,684, from 1972, the numbers began to decline [3]. In 2018, production reached 388,902, which is just 24% of the peak of 1971 [4]. The same is true for the obi. If the shipment value index is set at 100 for the peak, which came in 1975, then the index has dropped to 33.4 by 2014 [5]. These numbers indicate how severe the crisis is in the kimono industry. Kimonos come with a lot of strict rules. When you wear a kimono, you first have to think about status. That status includes wearing the proper kimono for a special celebration or an official event (the first formal ranking), which would be a black tomesode, or a semi-formal kimono called a homongi for women invited to a wedding, etc. Komon kimonos and yukatas are commonly worn. Obis also have status, and one must match the status of the obi and the kimono. A double-woven obi (fukuro-obi) goes with the black tomesode or the semi-formal homongi, and the Nagoya obi and share-tai obi go with a komon kimono. Also, the type of material you should wear changes with the seasons. In the height of summer, for example, you should wear cool materials such as hemp. In winter, 100% silk, which is a relatively warm material, should be worn [6]. Kimonos have different designs and patterns depending on the season and the event. For example, stripes or regular hexagons connected at the top, bottom, left, and right in a turtle-shell pattern may be worn all year round. However, designs with the morning glory, which blooms in summer, should only be worn in summer [7]. One has to think about the color combinations of the kimono and obi as well as the matching material. There are many elements to consider. In the era when the kimono was everyday wear, the parents would teach their children the rules, and the children would teach the grandchildren. But in an age when 69.6% of the women of working-age population work (2019) [8], there is no time for families to teach the rules, and it is difficult to know if one's obi and kimono match. Today, when people rarely wear kimonos and they are used less and less, the opportunity to learn these rules becomes even rarer. This is one of the reasons that Japanese people have fewer and fewer chances to wear kimono. That's why this research has focused on something that has been garnering interest in recent years, the CNN (Convolutional Neural Network) to apply machine learning to see whether a kimono and obi match or not, automatizing the process and improving the rate of correct answers, and we want to provide an environment where Japanese can wear kimonos whenever they want. Therefore, the aims of our research are to create the training data necessary to develop a system that can match kimono and obi by using CNN and deep learning technology, and to construct a network model which will learn by using that data. In order to consider the optimal model, we will compare a simple CNN model with the model using VGG16 which has high performance in image recognition.

In this paper, we will first introduce the research trends in digital fashion with reference to some literature, and the novelty and usefulness of this research. Next, we will go on to describe the method for creating training data for judging the quality of arrangement of Japanese kimono and obi using a simple CNN or VGG16 model. Then, we will present the results of performing validation using a model trained by these data and comparing the correct answer rates, and finally our conclusion.

II. Related Works

Various fashion-related studies using artificial intelligence technology have been made [9]. For example, in [10], this study examined how semantic information extracted from clothes images using computer vision could be used to improve the user experience in online shopping. There are studies that can recommend an item set instead of one type of item. In order to recommend a set of fashion items, a tensor decomposition approach is utilized [11]. Heterogeneous graphs linking fashion items make up stylish outfits and link items to their attributes [12]. Another example of using CNN is a study that extracts styles from Amazon image sets so that they can be recommended by style rather than item category [13]. Some studies have shown that using CNNs to extract image and shape features of items provided better recommendations than text-based ones [14, 15]. On the other hand, there is a digital archiving system for Japanese kimonos [16], but there is no system that recommends Japanese kimono with appropriate obi. As mentioned above, the arrangement of Japanese kimono and obi involves the tacit knowledge of kimono experts, and there is no research that makes this possible by CNN or deep learning.

III. Experiment

A. *Experiment using Images of Kimono and Obi to Determine Matches*

We prepared images of 100 homongi kimonos (images contributed by Kyoto Yuzen Corporative) and 20 images of double-woven obis (images contributed by Company Kyoto Kimono Ichiba). We had three people participate by deciding whether the 20 obis matched with the 100 kimonos. The participants were one man (in his 20s) and two women (in their 40s and 50s). Examples of the kimonos are shown in Fig. 1 and examples of the obis are shown in Fig. 2. The size of the images was 391×324 pix for the kimonos, and 391×261 pix for the obis displayed on a 32-inch monitor (NEC Multi Sync Lcd V323). The participants viewed the images from 50cm away at a vertical angle of about 42° and a horizontal angle of about 70°. Furthermore, the participants viewed the images in a dark room so that lighting would not affect their view.

Fig.3 shows the screen shot of the program that the participants used for evaluation of matching kimonos and obis, and Fig. 4 shows the experimental scene.

Fig. 1. Example of the Kimono Images Shown to Participants.

Fig. 2. Example of the Double-Woven Obis Shown to Participants.

Fig. 3. Screen shot of the Program that the Participants used for Evaluation of Matching Kimonos and Obis.

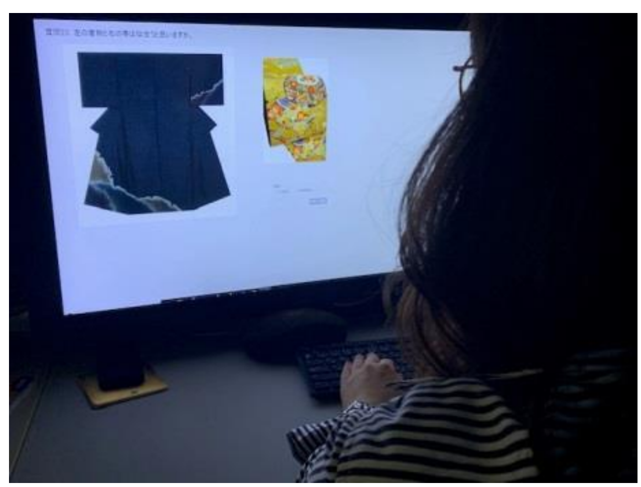


Fig. 4. Participants' view of Kimono and Obi on the Monitor.

The 100 kimono images and 20 obi images were combined into 2,000 combinations and input into the computer. The participants discussed whether the kimono and obi matched or not, and the majority opinion prevailed in this supervised learning exercise. Of the 2,000 combinations, 1,600 were chosen randomly as training data, and the remaining 400 were used as validation data. The 2,000 combinations divided into 1,600 used for training data and 400 used for validation data were randomly shuffled every time, and this process was repeated five times. The specs used this time are shown in Table 1. We used the Ubuntu 16.04 operating system and the Keras library. The GPU was GTX 1080 Ti. There were two different models for this research. The first used a simple CNN model (referred to below as CNN model). The other used the VGG16 model [17] (referred to below as VGG model). The VGG16 has a neural network that can finish learning about a convolution of more than 1 million images via the ImageNet database [18]. This network is 16 layers deep and is divided into objects by 1,000 (keyboard, mouse, pencil, types of animals, etc.).

B. *Experiment with the Simple CNN Model*

With the simple CNN model, we input 1,600 images of kimonos and 1,600 images of obis separately. The model structure that folded in the convolution of images in layer 1 (Conv2D) is shown in Fig. 5. The input was for both kimono images and obi images that were sorted into a convolution layer and a MaxPooling layer and applied to all bonded layers. They were bonded with the concatenate functions and sandwiched in with the other layers. Softmax was applied and a final output resulted. Furthermore, after each CNN layer and after the bonding of the 3rd layer with the other layers, the activation function ReLU was used in between the bonding of the third layer and the activation function and after the bonding of the 4th layer, Batch Normalization function was set. For the above model, Adam was used as an optimizer, and we used categorical cross-entropy for the loss function, setting the epoch number at 100 times. We repeated this series of process five times.

The format for each layer in Fig. 5 is shown in Table 2. Also, in Fig. 5, to be concise, we abbreviated the activation function to Batch Normalization.

TABLE. I. THE CALCULATORS USED

Calculator	DeepStation DK-1000 Personal Edition
Operating System	Ubuntu 16.04
Library	Keras
GPU	GTX1080 Ti

TABLE. II. FORMAT OF EACH LAYER

Input	Input([Vertical]×[Width]×[Channel])
Convolution Layer	Conv[kernel size]-[output]-[padding]
MaxPooling Layer	MaxPooling[pooling size]
Fully Bonded Layers	Dense[unit]
Dropout Layer	Dropout([Dropout ration])
Output	Output[No. of categories]

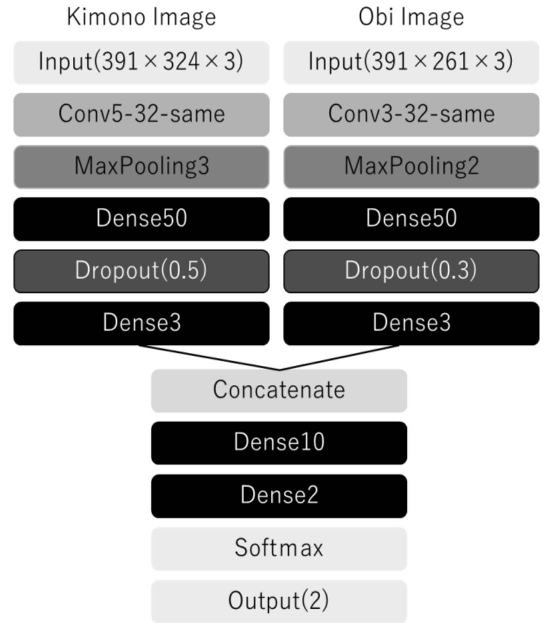


Fig. 5. Simple CNN Model Map.

C. *Experiment using VGG16 Model*

VGG is the model where the convolution layer and pooling layer are configured. VGG16 is when the convolution layer and other bonded layers are layered 16 deep. The unit model for VGG16 is shown in Fig. 6.

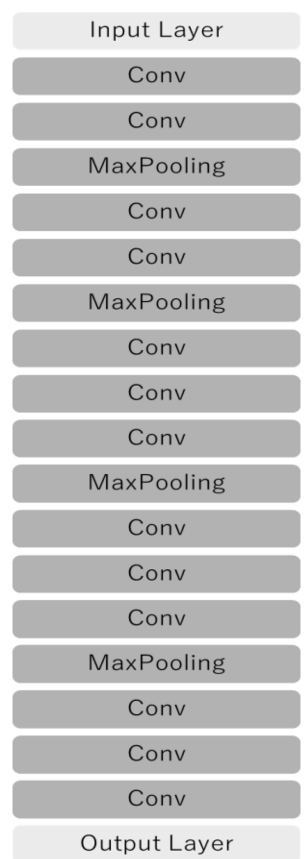


Fig. 6. Structure of VGG16.

As with the CNN model, a total of 2,000 combinations were used from the input data of 100 images of kimonos and the 20 images of obis. The participants discussed whether the kimono and obi matched or not, and the majority opinion prevailed in this supervised learning exercise. Of the 2,000 combinations, 1,600 were chosen randomly as training data, and the remaining 400 were used as validation data. The 2,000 combinations divided into 1,600 used for training data and 400 used for validation data were randomly shuffled every time, and this process was repeated five times.

Then, we input 1,600 images of kimonos and 1,600 images of obis separately and sent them to the VGG model. At this point, we used the learning convolution parameters of ImageNet. The learning went to the 15th and later layers, a process known as transfer learning. After that, all the layers were bonded with concatenate functions, Softmax was applied and a final output resulted. Furthermore, after the 3rd and 4th bonded layers were equipped with Batch Normalization, we used the activation function ReLU, which was set after the bonding of the 3rd layer and after Batch Normalization was set. SGD was used as an optimizer on the above model and the learning rate was set to 0.0001. We used categorical cross-entropy for the loss function. After a number of trials, about 50 epochs gave us the peak training results, so we set the epoch number to 50 times. We repeated this series of process five times. The formation of the different layers shown in Fig. 7 is the same as Table 2.

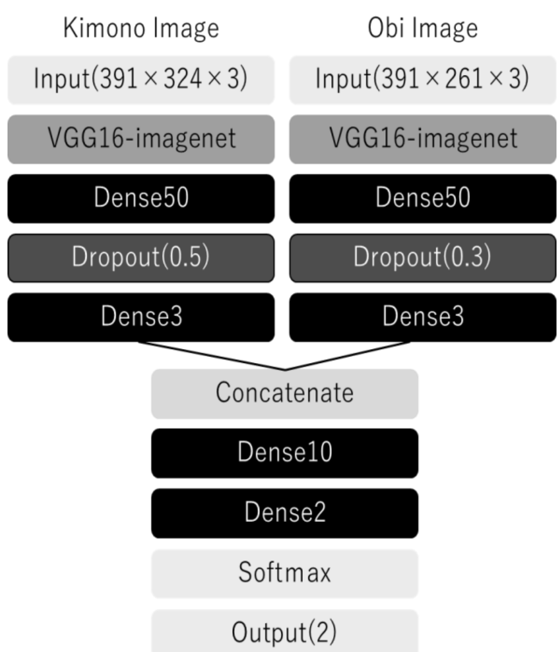


Fig. 7. Model using VGG16.

IV. Result

A. Results with the Simple CNN Model

Fig. 8 is a graph showing the transition of the accuracy rate when training data and validation data were used once. The results for the CNN model using just one layer unit is shown in Table 3. Using 100 epochs repeated five times, the average was 75.4%.

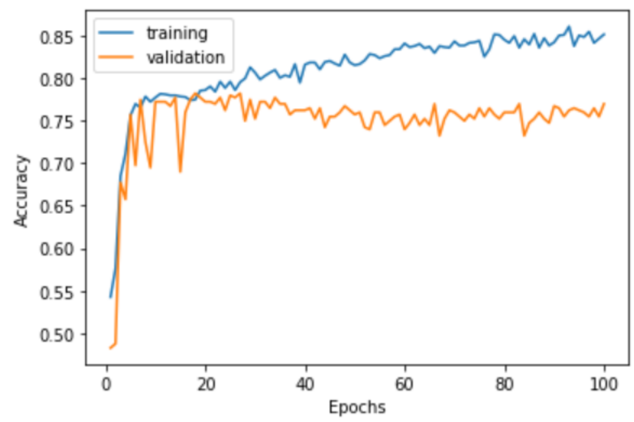


Fig. 8. Transition of Accuracy Rate for One Time of Training and Validation.

TABLE. III. Results from the Simple CNN Model

	Accuracy Rate
1 time	77.0%
2 times	74.3%
3 times	78.3%
4 times	70.0%
5 times	77.3%
Average	75.4%

B. Results from using the VGG Model

Fig. 9 shows the transition of the accuracy rate on the fourth attempt of five samples using training data and test data. The results from the VGG model are shown in Table 4. When 50 epochs are repeated 5 times, the average is 76%.

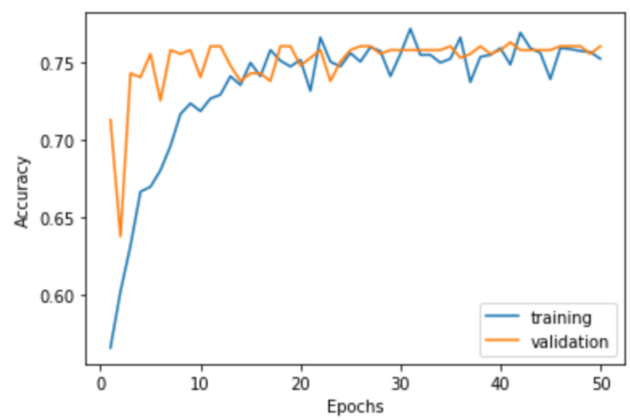


Fig. 9. Transition of the Accuracy Rate on the 4th Time using Training and Validation Data.

TABLE. IV. Results from the VGG16 Model

	Accuracy Rate
1 time	74.3%
2 times	78.0%
3 times	77.5%
4 times	76.0%
5 times	74.0%
Average	76.0%

V. Discussion

This time, we aimed to automate the arrangement of Japanese kimono and obi by using CNN and Deep Learning technologies. Then, we compared the average accuracy rates of the simple CNN model that has only one CNN layer and the model utilizing VGG model with each model repeating the epoch numbers five times using cross validation. The results showed that the simple CNN model averaged 75.4%, and the VGG16 model averaged 76.0%. The difference is 0.6%.

We hypothesized that the VGG model would produce far greater results, but that did not happen. This time, with a simple task of deciding whether a kimono and obi were a good match or not, the CNN model using just one layer averaged more than 75%, a perfectly acceptable result. It can be seen that there is no reason to use the many CNN layers available to VGG16. However, as shown in Fig. 7, there is a distinct gap in the CNN model between the accuracy rate of the training data and the validation data. This indicates the possibility of overtraining.

Consequently, we were able to research the accuracy rate when deciding whether obis matched certain kimonos, but deciding on the right kimono attire is not as simple as just whether the obi and kimono match. Actually, there are many rules for deciding which obis match which kimonos. The most important is the status of the kimono. In the long history of the kimono, this is something that started being said in 1976 [19]. Today, even though there are various styles for wearing a kimono and people express themselves freely with their clothing, conventional kimono rules cannot be ignored. Especially in public places and events, some people will continue to honor these rules. When we asked kimono experts about what to consider when matching kimonos and obis, they always said “status” first [20], but for average people who don’t have deep knowledge of the kimono, it’s clear that they focus only on color.

In other words, it became clear that they only focused on the compatibility of colors. In this research, when we created the supervised learning exercise, of our participants, the one man had no knowledge of the kimono, and he said in an interview that he only considered colors when deciding whether a kimono and obi go together. However, the two female participants said that they rejected his decision which had ignored the rules of matching kimono and obi. That said, this research showed that both the CNN model and the VGG model produced accuracy rates of higher than 75%, and these results might reflect not only the color combination but also other rules such as the status, texture, the seasons, and the meaning of the patterns of these items. This means that the tacit knowledge of the kimono experts can be implement by using these technologies.

VI. Conclusion

We attempted to build a system that automates the arrangement of Japanese kimono and obi using CNN and deep learning techniques. To this end, we created the original training data. In preparing the training data, considering the rules of matching kimono and obi, and training the model with the data, it was possible to determine whether or not the arrangement was appropriate with an accuracy of about 75% using both techniques.

As we move further in this research, we proceeded with our research only using the learning convolution parameters on the 15th and later layers based on the learning convolution parameters of the VGG16. Using VGG16 ImageNet as a foundation, we will look at whether the accuracy rate can be improved through learning when using kimono and obi images on all layers. There is also potential for the simple CNN model to over-train, so we would also like to expand our data and see if we can validate an improved accuracy rate.

The results of this research as related to Internet business shows that it would be possible to help customers with no kimono knowledge by recommending the right combinations. Also, there could be an application that examines a photo of your kimono and obi and tells whether they match or not, giving the kimono wearer more confidence. There are various ways to wear kimono, and people express themselves freely through their clothes. But even today, we can’t ignore the rules. Especially at public places and celebratory events, things will change a little as we go on, but traditions will be handed down. We hope that this research will help preserve the kimono culture.

Reference

[1] Kazuko Koizumi, Kimono of the showa era, KAWADE SHOBO SHINSHA, 2006 (in Japanese).

[2] Masuda Yoshiko, Japanese clothing history, Yoshikawa Kobunkan, 2010 (in Japanese).

[3] Kyoyuzen Komon Manufacturing Survey and Report, Kyoto Yuzen Cooperative, 2016 (in Japanese).

[4] Kyoyuzen Komon Manufacturing Survey and Report, Kyoto Yuzen Cooperative, 2019 (in Japanese).

[5] Outline of Nishijin machinery industry survey, The 21st Nishijin Machinery Industry Investigation Committee, 2016 (in Japanese).

[6] Analysis of the thought process for choosing suitable Kimono for Customer by the expert, Kumiko Komizo, Takashi Furukawa Noriaki Kuwahara, Kazunari Morimoto, LNCS, Vol. 9745, pp. 559-567, 2016.

[7] Kenzo Fujii, Kimono pattern, SEKAIBUNKA PUBLISHING INC., 2009 (in Japanese).

[8] White Paper on Gender Equality 2019, Gender Equality Bureau, Cabinet Office, 2019, ISBN-10: 4906955959 (in Japanese).

[9] Haosha Wang and Khaled Rasheed, Artificial intelligence in clothing fashion. In Proc. on the International Conference on Artificial Intelligence (ICAI), 2014.

[10] Zequn Li, Honglei Li, and Ling Shao, Improving Online Customer Shopping Experience with Computer Vision and Machine Learning Methods. In Proc. of International Conference on HCI in Business, Government and Organizations. pp.427-436, 2016.

[11] Yang Hu, Xi Yi, and Larry S Davis. Collaborative fashion recommendation: A functional tensor factorization approach. In Proc. of the 23rd ACM international conference on Multimedia, pp.129-138, 2015.

[12] Hanbit Lee and Sang-goo Lee, Style Recommendation for Fashion Items using Heterogeneous Information Network, In Proc. of Recommender Systems 2015.

[13] Qiang Liu, Shu Wu, and Liang Wang, DeepStyle: Learning User Preferences for Visual Recommendation. In Proc. of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.841-844, 2017.

[14] Satya Keerthi Gorripati and Anupama Angadi, Visual Based Fashion Clothes Recommendation with Convolutional Neural Networks, International Journal of Information Systems & Management Science, Vol. 1, No. 1, 2018.

[15] Ryosuke Saga and Yufeng Duan, Apparel Goods Recommender System Based on Image Shape Features Extracted by a CNN, *Proc.* of 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), DOI: 10.1109/SMC.2018.00406.

[16] Haruko Takahashi, An Image Digital Archive for Substantiating the Acculturation of Clothing Culture in Japan, *Proc.* of 2015 International Conference on Culture and Computing (Culture Computing), DOI: 10.1109/ Culture.and.Computing.2015.13.

[17] Simonyan, Karen and Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv: 1409. 1556, 2014.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In Proc. CVPR, pp. 2-9, 2009.

[19] Motoshige Ito and Yajima Takatoshi, Kimono culture and Japan, Nikkei Publishing Inc., 2016 (in Japanese).

[20] Choosing an Obi Suitable for Kimonos Eliciting a Sense of Hannari - From a color-based perspective -, Kumiko Komizo, Noriaki Kuwahara, The Journal of Ergonomic Technology,Vol.19, No.1, pp.45-56, 2019.

An Enhanced K-Nearest Neighbor Predictive Model through Metaheuristic Optimization

Allemar Jhone P. Delima[1]
College of Computing Education, Professional Schools
University of Mindanao, Matina, Davao City, Philippines

***Abstract*—The K-Nearest Neighbor (KNN) algorithm is vulnerable to noise, which is rooted in the dataset and has negative effects on its accuracy. Hence, various researchers employed variable minimization techniques before predicting the KNN in the quest so as to improve its predictive capability. The Genetic Algorithm (GA) is the most widely used metaheuristics for such purpose; however, the GA suffers a problem, which is its mating scheme bounded on its crossover operator. Thus, the use of the novel Inversed Bi-segmented Average Crossover (IBAX) was observed. In the present work, the crossover improved genetic algorithm (CIGAL) was instrumental in the enhancement of KNN's prediction accuracy. The use of the unmodified genetic algorithm had removed 13 variables; while the CIGAL then further removed 20 variables from the 30 total variables in the faculty evaluation dataset. Consequently, the integration of the CIGAL to the KNN (CIGAL-KNN) prediction model improved the KNN prediction accuracy to 95.53%. In contrast to the model of having the unmodified genetic algorithm (GA-KNN); the use of the lone KNN algorithm, the prediction accuracy is only at 89.94% and 87.15%, respectively. To validate the accuracy of the models, the use of the 10-folds cross-validation technique revealed a 93.13%, 89.27%, and 87.77% prediction accuracy of the CIGAL-KNN, GA-KNN, and KNN prediction models, respectively. The above results show that the CIGAL carried out an optimized GA performance and increased the accuracy of the KNN algorithm as a prediction model.**

Keywords—CIGAL-KNN; GA-KNN; IBAX operator; KNN algorithm; prediction models

I. Introduction

Prediction in research serves as a powerful tool in the process of planning that can provide the researcher with a likelihood of future events. Technically, this data mining [1-2] approach is driven by using experiences and applying statistical, mathematical, or computational methods [3-5].

For most of the organizations, prediction helps the administration in analyzing the data, which is needed for managerial decisions. The data can positively constitute an improvement in the excellence of organizations' services [4]. Various researches in the different areas employ several machine learning and data mining algorithms for the prediction that helps management in the decision and policy-making undertakings. Among the many predictive algorithms, the KNN is the simple, reliable, and one of the most commonly used algorithm for such prediction and classification purposes [6]. It has been used in crime mining [7], educational data mining [8], and healthcare services [9] and so on. However, the KNN algorithm is vulnerable to noise or irrelevant data features [10]. This problem is rooted in the dataset and leads to low prediction accuracy, resulting in a flawed prediction tantamount to unreliable decision making.

To address the above-mentioned problem, various researchers have employed the data reduction methods to optimize the number of variables within the dataset to improve KNN's predictive accuracy. A technique has been used to increase the prediction accuracy of KNN and through the integration of the genetic algorithm in the predictive model (GA-KNN) [11]. Among the many metaheuristics, the genetic algorithm is one of the most competent indexes used for global optimizations, feature selection, and data reduction algorithm that is widely used in the literature [12-14]. However, the accuracy result of the hybrid GA-KNN prediction model is still not good enough. The genetic algorithm still suffers a problem with premature convergence – a coupling-based problem bounds on the crossover operator of the GA [13]. To address the issue, one of the suggested solutions is to prevent the premature convergence to design an efficient crossover operator; thus, the creation of the novel Inversed Bi-segmented Average Crossover (IBAX) [15]. There is an upcoming need of developing a new crossover operator for the genetic algorithm, so as to carry out an optimized GA performance. And it is needed for variable minimization that will improve the accuracy of the KNN prediction model [16]. The use of predictive models in an organization with erroneous data will lead to flawed predictions, which perpetuates the risk of additional harm that negatively affects the decision of the management in an organization.

This study aimed to increase the accuracy of the KNN algorithm as a prediction model through the integration of the novel Inversed Bi-segmented Average Crossover (IBAX) – a new crossover operator of the genetic algorithm applied in the faculty evaluation data. Specifically, the study aimed to determine the reduction in the number of variables using the unmodified genetic algorithm as against the crossover-improved genetic algorithm proposed by [15] and calculate the improvement in the accuracy of KNN when integrated with the data reduction techniques.

II. Literature Review

The KNN algorithm being one of the most effective nonparametric techniques due to its simplicity, suffers problems in its accuracy and k sensitivity. Various researches were conducted to address the issue of k sensitivity. A robust generalized mean distance-based k-nearest neighbor classifier was proposed in the quest to prevent the degradation of KNN-based performance due to neighborhood k sensitivity. The GMDKNN observed the generalized mean distance as preliminary that measures the distance similarity of the sample query and the k nearest neighbor. In general, the GMDKNN introduced the multi-generalized mean distances and the nested generalized mean of each class. The proposed method can employ more nearest neighbors for the favorable classification and has less sensitiveness to the values of k [17]. Further, another technique using the local mean representation-based KNN algorithm (LMRKNN) was proposed to solve the repetitive problem of the KNN. In the LMRKNN, "the categorical k-nearest neighbors of a query sample are first chosen to calculate the corresponding categorical k-local mean vectors, and then the query sample is represented by the linear combination of the categorical k-local mean vectors; finally, the class-specific representation-based distances between the query sample and the categorical k-local mean vectors are adopted to determine the class of the query sample." The simulation results revealed that the LMRKNN model outperformed the other relative KNN-based methods used in the study [18].

Another modification on the KNN method was proposed employing the two locality constrained representation for k nearest neighbors. The method is called weighted representation-based k-nearest neighbor rule (WRKNN), while the other is termed as the weighted local mean representation-based k-nearest neighbor rule (WLMRKNN). In WRKNN, the linear combination of the k nearest neighbor from each class is represented as the test sample, and the localities of k-nearest neighbors per class as the weights constrain their corresponding representation coefficients. "Using the representation coefficients of k-nearest neighbors per class, the representation-based distance between the test sample and the class-specific k-nearest neighbors is calculated as the classification decision rule." For the latter, "the k-local mean vectors of k-nearest neighbors per class are first calculated and then used for representing the test sample. In the linear combination of the class-specific k-local mean vectors to represent the test sample, the localities of k-local mean vectors per class are considered as the weights to constrain the representation coefficients of k-local mean vectors. The representation coefficients are employed to design the classification decision rule which is the class-specific representation-based distance between the test sample and k-local mean vectors per class." The simulation results revealed that the proposed methods perform better with less k sensitivity as against the local mean-based k-nearest neighbor (LMKNN) [19], collaborative representation-based nearest neighbor (CRNN) [20], and multi-local means-based nearest neighbor (MLMNN) [21-22].

Extent on the nearest neighbor query, there have been numerous researches that were conducted. To name some, a novel data structure through buffer kd-tree for processing massive nearest neighbor queries on GPUs was introduced [23]. Similarly, the performance of the KNN algorithm was improved based on the revised buffer kd-tree integration. A fast neighbor search through the revised kd-tree was realized although the method is not suitable for high dimensional data [24]. With respect to high dimensional data problems, the scalable nearest neighbor method through the introduction of the k-means tree for fast approximate matching of binary features along with the k-d forest was proposed and found to be effective in addressing the issues arising when scaling to very large size data sets [25]. Other proposed studies include the implementation of the fast k-nearest neighbor search via dynamic continuous indexing (DCI) [26], prioritized dynamic continuous indexing (PDCI) [27], and a fast exact nearest neighbor search algorithm based on semi-convex hull tree over large scale data all in the quest to find the k-nearest neighbor objects to a given point in class and space [28].

Premised on improving the prediction rate of the KNN algorithm due to its known repetitive problem, various researchers employed data reduction using optimization algorithms for such purpose. The use of PSO and CFS were instrumental for feature selection in the quest to improve KNN's predictive accuracy in predicting the occurrence of malignant tumors in the skeletal bones called sarcoma. The simulation results revealed that the PSO-KNN method attained an 85% accuracy compared to the CFS-KNN model with 81% accuracy [29].

Further, the PSO-KNN model and CART algorithm, when used in the water level estimation and water quality forecast in Poyang Lake in China, revealed a prediction accuracy of 86.68% using the PSO-KNN model that outperformed the average prediction of the CART algorithm with 81.76% prediction accuracy. [30]. Another effective technique to increase the accuracy of the KNN is the use of principal component analysis (PCA) before the prediction. The PCA-KNN hybrid model yielded an accuracy of 61.34%. The PCA-KNN model outperformed other predictive models in heart disease diagnosis using the heart disease dataset from the UCI machine learning repository [31].

The introduction of the gravitational search algorithm (GSA) serves as a feature reduction technique to increase the KNN's prediction accuracy for disease prediction using biomedical data. Using the GSA on heart dataset, there was a 64.61% reduction in the features while 57.64% and 77.77% of the features were removed using the dermatology and breast cancer datasets, respectively. In general, an average of 66% of the removed variables considerably improved the prediction accuracy of the KNN algorithm, increasing the accuracy from 64.81% to 82.96% for heart dataset to name one [32].

The genetic algorithm as a feature selection technique enhanced the performance of the KNN and Naïve Bayes algorithms in diabetes detection. The GA-KNN model attained an accuracy of 83.12%, while 81.82% prediction accuracy was depicted using the GA-NB model, making the GA-KNN as the optimal model for prediction of the healthy or diabetic patient [11]. Furthermore, the use of GA-KNN and GA-SVM models was observed in disease diagnosis using gene expression levels. The GA-KNN model obtained a 92.68% prediction accuracy on the prostate dataset [33].

III. Methodology

The enhanced prediction model is composed of two significant stages, the variable optimization stage, and the prediction stage. The details are presented in Fig. 1.

The study enhanced the KNN predictive model by integrating the crossover-improved genetic algorithm having the IBAX operator. The use of the CIGAL was instrumental in the pre-processing step in optimizing variables within the dataset. The identified significant variables through the generation of the enhanced GA, which is the CIGAL were used as input for the prediction stage in the quest to obtain a better prediction accuracy of the model.

A. Data

The responses of 597 random student-respondents from the four State Universities and Colleges (SUC) in Caraga Region, Philippines namely; the Surigao State College of Technology (SSCT), Caraga State University (CSU), Surigao del Sur State University (SDSSU), and Agusan Sur State College of Agriculture and Technology (ASSCAT) in the evaluation of the faculty instructional performance for the 2nd semester of S.Y. 2016-2017 served as datasets of the study. Out of the total number of records, 70% and 30% data composition for training and testing were used for prediction. There are thirty variables that represented the instructional performance of the faculty, having divided into six (6) parts as to wit: methodology, classroom management, student discipline, assessment of learning, student-teacher relationship, and peer relationship as depicted in Table 1.

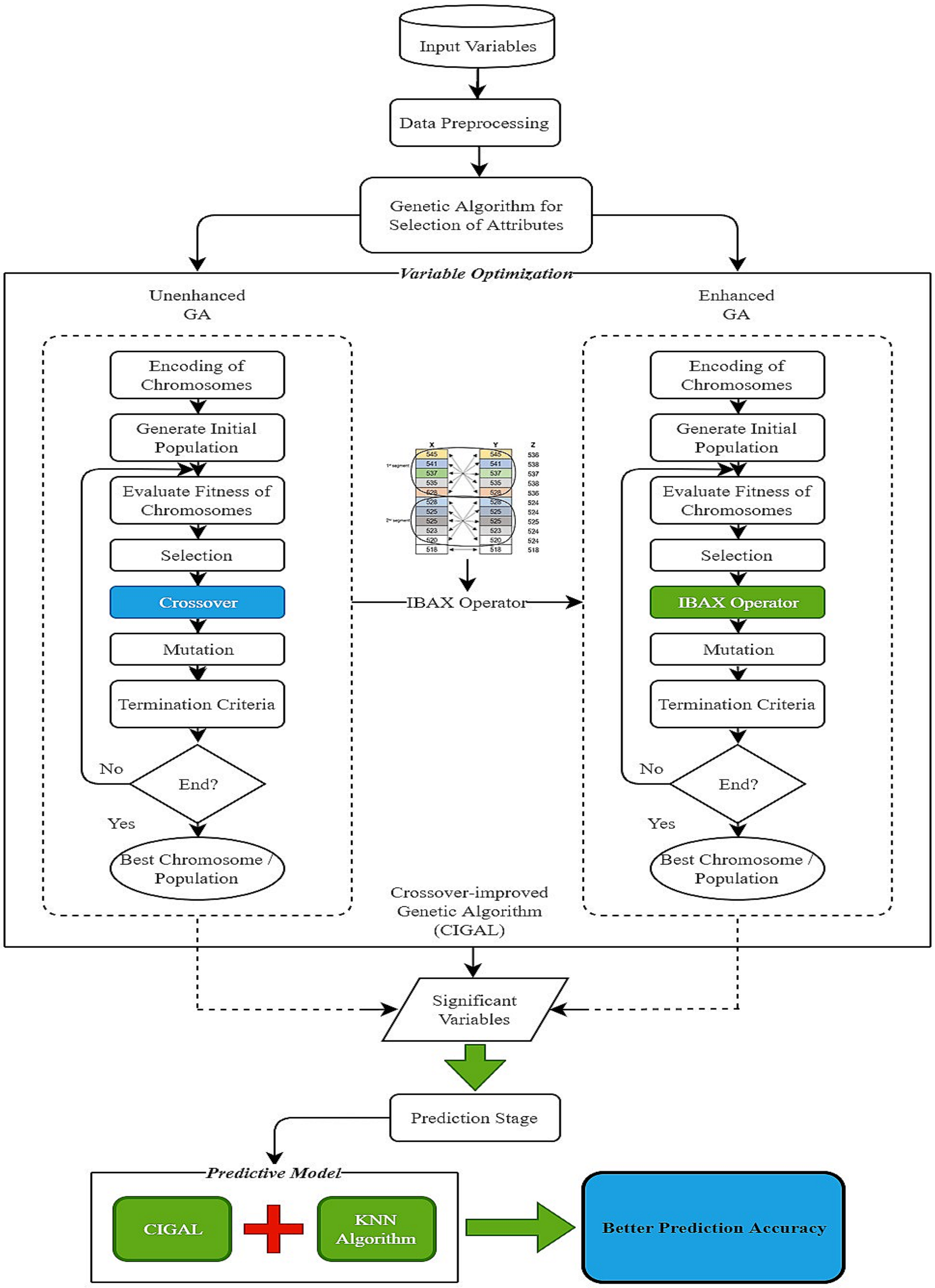


Fig. 1. Conceptual Framework of the Study.

TABLE. I. VARIABLES USED IN THE STUDY

Category	Description As a student, I have observed that my instructor/professor	Variable
Methodology	Utilizes varied designs/ techniques/ activities suited to the different types of learners.	M1
	Explains learning goals and instructional procedures to the students.	M2
	Uses real-life examples in the class to sustain the student's interest in learning.	M3
	Creates a situation that encourages students to use critical thinking.	M4
	Delivers accurate/relevant/updated content knowledge.	M5
Classroom Management	Establishes routines to maximize instructional time.	C1
	Organizes and assigns the daily cleaners.	C2
	Employs an effective system of the classroom set-up.	C3
	Employs strategies to maximize the use of resources in learning activities.	C4
	Implements rules/policies inside the classroom.	C5
Student Discipline	Handles behavior problems concerning the student's rights.	SD1
	Imposes disciplinary sanction(s) to the misbehaving student(s).	SD2
	Encourages students to submit requirements on time.	SD3
	Motivates students to respect each other.	SD4
	Allows students to exercise their creativity.	SD5
Assessment of Learning	Constructs a valid and reliable formative and summative test.	A1
	Uses appropriate non-traditional assessment techniques and tools (i.e. portfolio, journals, rubric, etc.)	A2
	Interprets and uses test results to improve teaching and learning.	A3
	Uses tools for assessing authentic learning.	A4
	Provides timely and accurate feedback to students.	A5
Student-teacher relationship	Encourages students to participate in class/school activities actively.	ST1
	Allows students to communicate directly to him/her.	ST2
	Provides equal opportunities for all students.	ST3
	Promotes teamwork among students.	ST4
	Makes him/herself available to students.	ST5
Peer relationship	Demonstrates appropriate behavior in dealing with students/peers/superiors.	P1
	Manifests flexibility, when deemed necessary.	P2
	Exhibits collegiality with colleagues.	P3
	Observes professionalism at all times.	P4
	Empathizes other needs and concerns.	P5

B. *Modification on the Crossover Operator of Genetic Algorithm*

In this study, the use of the new crossover operator of GA called inversed bi-segmented average crossover developed by [15], which is the modification of the traditional average crossover operator, was instrumental. The existing genetic algorithm operator called average crossover chooses the first gene of the first and second chromosomes. An offspring is produced by calculating the average of the mated genes. The process is performed repeatedly until the last genes of the two chromosomes have produced its offspring. The detailed flow of the average crossover mechanism is presented in Table 2.

Meanwhile, the inversed bi-segmented average crossover operator works by segmenting the chromosomes (x and y) into two and inversely computing the average of genes within each segment created. For each segment, the process entails a repeated performance until the last gene of the first chromosome mates with the first gene of the second chromosome. The detailed flow of the new crossover mechanism is presented in Table 3, and the graphical representation of the IBAX operator is shown in Fig. 2.

C. *Variable Optimization*

The process identified the significant variables in the dataset which are instrumental as input on the KNN algorithm for the prediction stage in the quest to improve its prediction accuracy. The number of variables in the dataset is optimized with the help of the CIGAL having the IBAX mating scheme. The complete flow of the improved GA with the IBAX mating scheme is shown in Table 4.

TABLE. II. THE FLOW OF THE AVERAGE CROSSOVER OPERATOR OF THE GENETIC ALGORITHM

Step No.	Steps
1	Select the first gene of the first chromosome (X) and the first gene of the second chromosome (Y).
2	Create one offspring (Z) out of the two genes selected using the average formula
3	$Z = [X + Y] / 2$
4	Repeat until the last gene of the first and second chromosomes are mated and have produced offspring.

TABLE. III. THE FLOW OF THE IBAX OPERATOR OF THE GENETIC ALGORITHM

Step No.	Steps
1	Count the number of genes found in the chromosomes. Identify if the variables are in odd or even count.
2	Segment the chromosomes (x and y) by dividing the total number of genes in the chromosomes into two. Make sure that both the first and second segments must contain an equal number of genes.
3	In the first segment, create offspring Z for each gene by inversely pairing the first gene from chromosome X to the last gene on chromosome Y. Repeat until the last gene of the chromosome X and the first gene of the chromosome Y have inversely mated and have produced an offspring using the average formula.
4	$Z = [X + Y] / 2$
5	Execute the same process on the second segment until genes from all segments have produced offspring. In the case of odd datasets, the last genes of the chromosomes will not be combined in the second segment and will automatically be mated with each other to produce offspring.

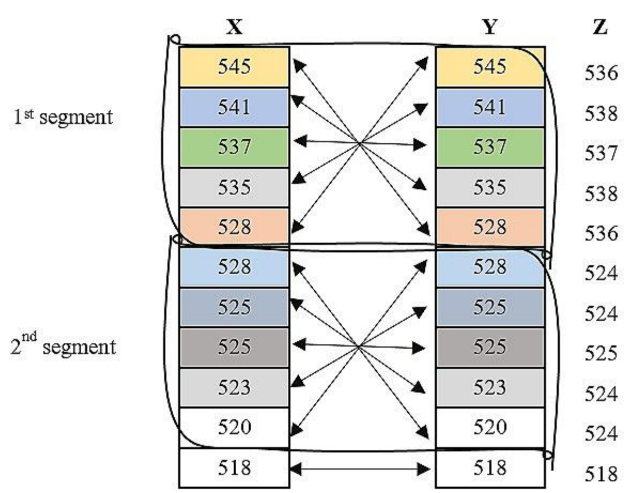


Fig. 2. Graphical Representation of the IBAX Operator.

To further test the effectiveness and the reduction rate of the CIGAL having the IBAX operator, a comparison on the reduction using the unmodified genetic algorithm having the roulette wheel selection function, original average crossover as the crossover operator, and the swapping mutation function, were instrumental. For the CIGAL, the rank-based selection function, the inversed bi-segmented average crossover operator, and the swapping mutation function were also utilized, executing both genetic algorithms for ten generations. In each generation, the new fitness value is calculated based on the result of the crossover function. The variables with the lowest fitness value after each generation for ten generations are aptly removed.

D. Enhanced Predictive Model

The significant variables determined by the crossover-improved genetic algorithm having the IBAX operator and the unmodified genetic algorithm with the original average crossover were instrumental in the prediction using the KNN algorithm having a *k* value of 3. The simulation of the KNN employed the Waikato Environment for Knowledge Analysis (WEKA) software with version 3.8.2. The detailed flow of the K-Nearest Neighbor algorithm is shown in Table 5.

TABLE. IV. CROSSOVER-IMPROVED GENETIC ALGORITHM INTEGRATING THE IBAX OPERATOR

Step No.	**Steps**
1	Specify the number of chromosomes and generations, as well as the value of crossover and mutation rates
2	Generate an initial chromosome-chromosome number of the population and the initialization of values of the genes based on the variables of an effective faculty instructional performance and calculate its fitness function
3	Evaluation of fitness value of chromosomes by calculating the objective function (Process steps 3-6 until the number of generations is met)
4	The use of rank-based selection function
5	Crossover having the IBAX operator
6	Mutation
7	Solution (Best Chromosomes)

TABLE. V. KNN ALGORITHM

Step No.	**Steps**
1	For a training set $A = \{(a_1, b_1), ..., (a_T, b_T)\}$, the n^{th} training sample is represented by $a_n \in A$, and $b_n \in \{w_1, w_2, ..., w_c\}$ represents the class label of the n^{th} training sample; and the total number of samples in the training set is represented by T where the total number of classes is the c.
2	Assign k value of 3
3	**for all** (Training samples $(n = 1, 2, ..., T)$) **do**
4	Calculate the distance between the testing sample (a_{test}) and the training samples (a_n) as follows: $d_n = \sum_{n=1}^{T} (a_n - a_{test})^2$.
5	**end for**
6	The nearest *k*-training samples will be selected, such as the minimum k distances.
7	Assign the class which has the most samples among the k-nearest samples to the testing sample.

E. Prediction Accuracy Evaluation

The allocation of 70% and 30% data composition for training and testing determined the accuracy of the KNN prediction model when integrated with the CIGAL having the IBAX operator (CIGAL-KNN) as against the model when integrated with the unmodified genetic algorithm (GA-KNN). The predictive capability of the lone KNN algorithm as tested, further calculated the degree of improvement in the accuracy of the data reduction techniques.

F. Model Validation

The use of 10 folds cross-validation scheme was instrumental in validating the accuracy of the model. To select the optimal model for prediction, one must produce the highest accuracy rate with a lower statistical error value of the root mean squared error (RMSE). Along with it, the use of mean absolute error (MAE), precision, recall, and F-measure conformed to the formula as derived from the study of [34] as to wit:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

$$MAE = \sum_{t=T+1}^{T+h} | \widehat{y}_t - \widehat{y}_t | / h \tag{5}$$

$$RMSE = \sqrt{\sum_{t=T+1}^{T+h} (\widehat{y}_t - \widehat{y}_t)^2 / h} \tag{6}$$

where TP, TN, FP, and FN mean true positive, true negative, false positive and false negative, respectively. In the Eqs. (5) and (6), the forecast sample is represented by $T+1,...,T+h$ whereas the actual value is denoted by y_t along with $\widehat{y}_t$ representing the forecasted value in the period t. The lower the statistical error value is, the better the forecasting ability of the model.

IV. Results and Discussion

A. Variable Minimization Results

The variable minimization results utilized the CIGAL and compared the generated results using the unmodified GA. The index result is shown in Table 6.

The simulation results show that the crossover-improved genetic algorithm having the inversed bi-segmented average crossover outperformed the unmodified genetic algorithm in reducing the number of variables in the dataset. The use of the unmodified genetic algorithm has removed 13 of the variables from the dataset. From the 30 total number of variables, it was reduced up to 17. Meanwhile, the crossover-improved genetic algorithm having the inversed bi-segmented average crossover further reduced the number of variables in the dataset leaving only ten variables in general. The indexed simulation result generated by the novel crossover operator of the genetic algorithm is shown in Table 7.

In the first generation, the variables C2 and C3 obtained the lowest fitness value among the variables in the group, removing it from the chromosomes. The removed variables will have no chance to be included for the next generation. From the 30 variables, it was reduced to 28 from the first generation alone. The succeeding generations have removed variables ST2 and C1 for the second generation, variables ST4 and A2 for the third, variables P2 and A1 for the fourth generation, variables ST3 and A4 for the fifth generation, variables C5 and ST5 for the sixth generation, variables SD5 and SD2 for the seventh generation, variables A5 and M1 for the eighth generation, variables A3 and P3 for the ninth generation and lastly, the variables P4 and M5 were removed on the tenth generation. From the 30 total variables, the CIGAL removed 20 variables from the dataset and retained variable M2, M3, M4, C4, SD1, SD3, SD4, ST1, P1, and P5 as these are identified to be instrumental for prediction. The graphical representation of the variable minimization using the genetic algorithms is shown in Fig. 3.

TABLE. VI. Variable Minimization Result

Genetic Algorithms	Number of Variables	Number of Variables Removed	Number of Variables Left
Unmodified GA	30	13	17
CIGAL-IBAX	30	20	10

TABLE. VII. Indexed Simulation Result using the CIGAL

Number of Generations	Number of Variables Left	Number of Variables Removed	Variables Removed	Percentage
0	30	0	-	-
1	30	2	C2, C3	6.66%
2	28	2	ST2, C1	6.66%
3	26	2	ST4, A2	6.66%
4	24	2	P2, A1	6.66%
5	22	2	ST3, A4	6.66%
6	20	2	C5, ST5	6.66%
7	18	2	SD5, SD2	6.66%
8	16	2	A5, M1	6.66%
9	14	2	A3, P3	6.66%
10	12	2	P4, M5	6.66%
10	-	-	-	-
Total Percentage of Variables Removed				66.66%

The CIGAL with the IBAX operator removed 66.66% of the variables from the dataset as shown in Fig. 3. The simulation results revealed that the amount of reduction varies to the genetic algorithm used. The notion of dropping one or more variables within the dataset in the quest to help reduce dimensionality is certain. Therefore, the removal of 66.66% of the variables is acceptable since the 60% ratio of feature reduction is suitable, as orchestrated by the work of [35].

B. Prediction Model Accuracy Evaluation

The 70 and 30 percent data composition for training and testing was observed to evaluate the accuracy of the prediction models performed in WEKA software. The comparative results of the prediction model with the crossover-improved genetic algorithm (CIGAL-KNN) as against the prediction model with the unmodified genetic algorithm (GA-KNN) is shown in Table 8. The predictive capability of the KNN algorithm was also tested without the variable reduction stage.

The simulation results showed that there was an increase in the accuracy of the models with the integration of GA, particularly with the CIGAL. The CIGAL-KNN prediction model outperformed the GA-KNN model and the model having the KNN algorithm alone with 95.53%, 89.94%, and 87.15% correctly classified instances, respectively. Using the 70% and 30% data composition, the optimal model for predicting the accuracy of the responses on the faculty instructional performance evaluation in the four SUCs in the Caraga Region is the model with CIGAL having the IBAX operator integrated to the KNN. Fig. 4 shows the graphical representation of the accuracies obtained using the three prediction models.

C. Prediction Model Validation Results

The predictive capability of the models was validated using the ten folds cross-validation scheme performed in WEKA. The accuracy results of the model using the validation scheme are shown in Table 9.

Based on the simulated validation result of the predictive models in Table 9, it can be seen that the CIGAL-KNN model still obtained the highest correctly classified instances of 93.13% as against the 89.27% and 87.77% prediction accuracies for GA-KNN and KNN predictive models. The performance evaluation of the models made use of the RMSE and MAE forecast statistical error tools. Both statistical error tests revealed a zero-based value for RMSE and MAE with the lowest error value of 0.21 and 0.07, respectively, for the CIGAL-KNN model. The low estimated values revealed how concentrated the prediction is using the crossover-improved genetic algorithm-based KNN prediction model. A low statistical error value depicts an ideal and desirable model for a good forecast.

To further evaluate and compare the performance of the three models, the precision, recall, and F-measure performance metrics estimation was carried out. An overall precision score of 93.1% was depicted in the CIGAL-KNN model. The precision of the model in testing the accurateness of those predicted positives from the responses in the evaluation data is optimal. The recall metric denotes that 93% of the tested instances within the dataset were retrieved by the model correctly and the F-measure determines the 93% balance performance of the model. The graphical representation of the model validation in terms of accuracy is shown in Fig. 5.

TABLE. VIII. INDEXED COMPARATIVE ACCURACY RESULTS OF KNN, GA-KNN, AND CIGAL-KNN PREDICTIVE MODELS

Predictive Model	Accuracy
KNN Algorithm	87.1508%
GA-KNN	89.9441%
CIGAL-KNN	95.5307%

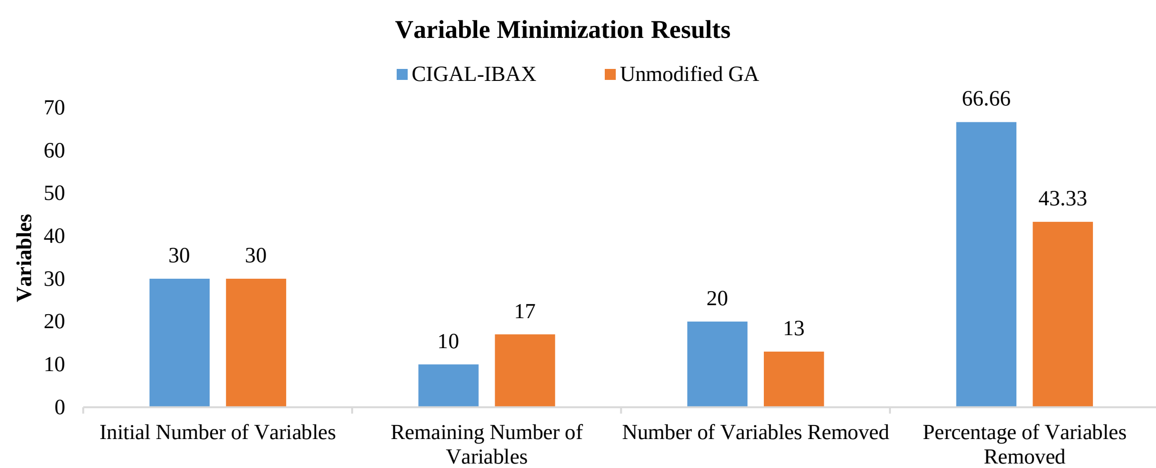


Fig. 3. Comparative Result for Variable Minimization using Genetic Algorithms.

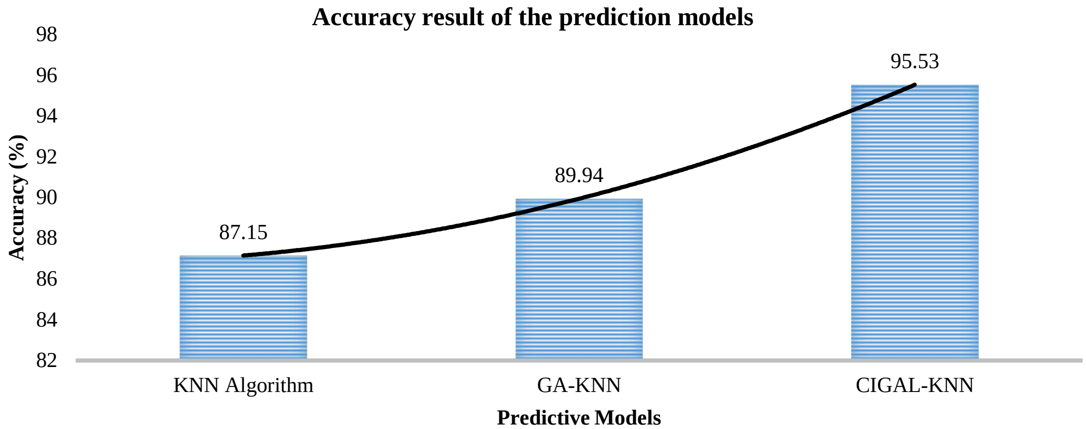


Fig. 4. Graphical Representation of the Accuracy Results of the Prediction Models.

TABLE. IX. INDEXED VALIDATION RESULT OF THE PREDICTIVE MODELS

Predictive Models	Accuracy	RMSE	MAE	Recall	Precision	F- Measure
KNN Algorithm	87.7722%	0.2981	0.1385	0.878	0.876	0.877
GA-KNN	89.2797%	0.2595	0.1119	0.893	0.892	0.892
CIGAL-KNN	93.1323%	0.2146	0.0786	0.931	0.932	0.930

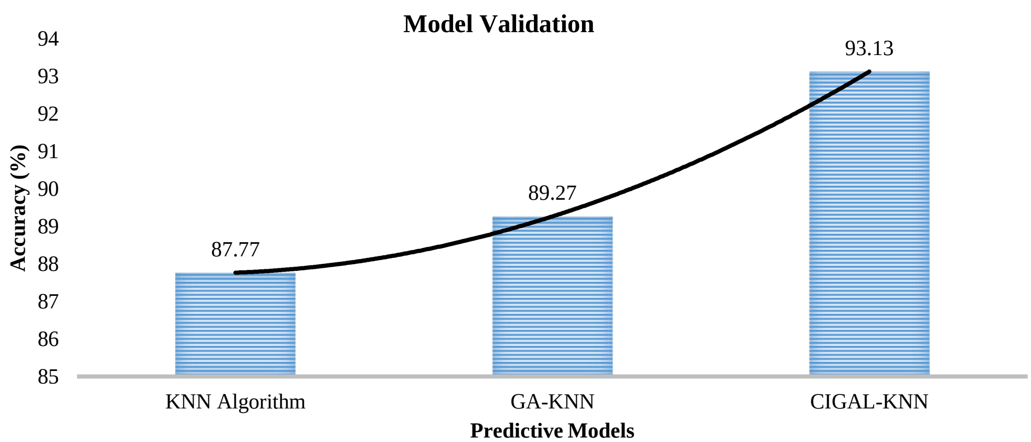


Fig. 5. Graphical Representation of the Model Validation Results.

V. CONCLUSION

The integration of the novel Inversed Bi-segmented Average Crossover to the genetic algorithm has paved the way for a more enhanced GA when it comes to optimization problems. Consequently, the incorporation of the crossover-improved genetic algorithm to the KNN has led to an increase in the accuracy of a prediction model. Since prediction accuracy affects the decision of the management of an organization, increasing the accuracy of prediction models is viewed as necessary. The variable minimization using the unmodified genetic algorithm removed 13 of the variables from the 30 total variables in the dataset leaving 17 variables to be used for prediction. Meanwhile, the crossover-improved genetic algorithm having the novel IBAX operator outperformed the minimization capabilities of the unmodified GA, further removing 20 variables from the 30 total variables in the dataset, leaving ten variables for prediction. In general, the integration of the crossover-improved genetic algorithm to the KNN predictive model (CIGAL-KNN) yielded an increase in prediction having a 95.53% accuracy against the identified 89.94% prediction accuracy with the integration of the unmodified GA (GA-KNN), and 87.15% prediction accuracy utilizing the lone KNN algorithm.

Premised on the conclusions of the study, the use of other variable minimization, feature selection, and global optimization algorithms aside from the GA is recommended in the continued quest to improve the accuracy of the KNN predictive model and present a comparative analysis of the results. The utilization of the CIGAL on other predictive algorithms aside from the KNN is recommended to come up with a comprehensive literature review on the latest modifications to the different prediction models. Further, future researchers are encouraged to utilize the prediction model in other real-life datasets.

References

[1] J. P. Delima, "Predicting scholarship grants using data mining techniques," International Journal of Machine Learning and Computing, vol. 9, no. 4, pp. 513-519, 2019.

[2] A. J. P. Delima, "Applying data mining techniques in predicting index and non-index crimes," International Journal of Machine Learning and Computing, vol. 9, no. 4, pp. 533-538, 2019.

[3] M. J. Rezaee, M. Jozmaleki, and M. Valipour, "Integrating dynamic fuzzy C-means, data envelopment analysis and artificial neural network to online prediction performance of companies in stock exchange," Physica A, vol. 489, pp. 78-93, 2018.

[4] U. O. Cagas, A. J. P. Delima, and T. L. Toledo, "PreFIC: Predictability of faculty instructional performance through hybrid prediction model," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 7, pp. 22-25, 2019.

[5] A. J. P. Delima and M. T. Q. Lumintac, "Application of time series analysis for philippines' inflation prediction," International Journal of Recent Technology and Engineering, vol. 8, no. 1, pp. 1761-1765, 2019.

[6] S. Fei, "The hybrid method of vmd-psr-svd and improved binary pso-knn for fault diagnosis of bearing," Shock and Vibration, vol. 2019, pp. 1-7, 2019.

[7] V. Vishnupriya and M. Valarmathi, "An effective data mining techniques for analyzing crime patterns," IOSR Journal of Computing Engineering, vol. 1, pp. 26-30, 2017.

[8] M. Kumar, A. J. Singh, and D. Handa, "Literature survey on student's performance prediction in education using data mining techniques," International Journal of Education and Management Engineering, vol. 7, no. 6, pp. 40-49, 2017.

[9] A. Rairikar, V. Kulkarni, V. Sabale, H. Kale, and A. Lamgunde, "Heart disease prediction using data mining techniques," in International Conference on Intelligent Computing and Control, I2C2, 2017, pp. 1-8.

[10] D. García-gil, J. Luengo, S. García, and F. Herrera, "Enabling smart data : noise filtering in big data classification," Information Sciences Journal, vol. 479, pp. 135-152, April 2019.

[11] R. N. Patil, and S. C. Tamane, "Upgrading the performance of knn and naïve bayes in diabetes detection with genetic algorithm for feature selection," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 3, no. 1, pp. 1371-1381, 2018.

[12] A. J. P. Delima, "An experimental comparison of hybrid modified genetic algorithm-based prediction models," International Journal of Recent Technology and Engineering, vol. 8, no. 1, pp. 1756-1760, 2019.

[13] M. Y. Orong, A. M. Sison, and R. P. Medina, "A hybrid prediction model integrating a modified genetic algorithm to k-means segmentation and c4.5," in TENCON 2018 - 2018 IEEE Region 10 Conference, October 2018, pp. 1853-1858.

[14] M. Mafarja, I. Aljarah, A. A. Heidari, A. I. Hammouri, H. Faris, and A. M. Al-zoubi, "Evolutionary population dynamics and grasshopper optimization approaches for feature selection problems," Knowledge-Based Systems, vol. 145, no. 1, pp. 25-45, April 2018.

[15] A. J. P. Delima, A. M. Sison, and R. P. Medina, "A modified genetic algorithm with a new crossover mating scheme," Indonesian Journal of Electrical Engineering and Informatics, vol. 7, no. 2, pp. 165-181, July 2019.

[16] A. J. P. Delima, A. M. Sison, and R. P. Medina, "Variable reduction-based prediction through modified genetic algorithm," International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 356-363, 2019.

[17] J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao, and H. Yang, "A generalized mean distance-based k-nearest neighbor classifier," Expert Systems with Applications, vol. 115, pp. 356-372, 2019.

[18] J. Gou, W. Qiu, Z. Yi, Y. Xu, Q. Mao, and Y. Zhan, "A local mean representation-based k-nearest neighbor classifier," in ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 3, pp. 29:1-29:5, May 2019.

[19] Y. Mitani and Y. Hamamoto, "A local mean-based nonparametric classifier," Pattern Recognition Letters, vol. 27, no. 10, pp. 1151-1159, July 2006.

[20] W. Li, Q. Du, F. Zhang, and W. Hu, "Collaborative-representation-based nearest neighbor classifier for hyperspectral imagery," IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 2, pp. 389-393, February 2015.

[21] J. Gou, W. Qiu, Q. Mao, Y. Zhan, X. Shen, and Y. Rao, "A multi-local means based nearest neighbor classifier," Proc. 2017 IEEE 29th International Conference on Tools with Artificial Intelligence, June 2018, pp. 448-452.

[22] J. Gou, W. Qiu, Z. Yi, X. Shen, Y. Zhan, and W. Ou, "Locality constrained representation-based k-nearest neighbor classification," Knowledge-Based Systems, vol. 167, pp. 38-52, March 2019.

[23] F. Gieseke, J. Heinermann, C. Oancea, and C. Igel, "Buffer k-d trees: Processing massive nearest neighbor queries on GPUs," Proc. 31st International Conference on Machine Learning, ICML 2014, January 2014, pp. 172-180.

[24] Y. Chen et al., "Fast neighbor search by using revised k-d tree," Information Sciences, vol. 472, pp. 145-162, January 2019.

[25] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 11, pp. 2227-2240, 2014.

[26] K. Li and J. Malik, "Fast k-nearest neighbour search via dynamic continuous indexing," Proc. 33rd International Conference on Machine Learning, ICML 2016, June 2016, pp. 671-679.

[27] K. Li and J. Malik, "Fast k-nearest neighbour search via prioritized DCI," Proc. ICML'17 Proceedings of the 34th International Conference on Machine Learning, ICML 2017, 2017, vol. 70, pp. 2081-2090.

[28] Y. Chen et al., "Semi-convex hull tree: Fast nearest neighbor queries for large scale data on GPUs," Proc 2018 IEEE International Conference on Data Mining, ICDM, 2018, pp. 911-916.

[29] K. Baskaran, R. Malathi, and P. Thirusakthimurugan, "Feature fusion for FDG-PET and MRI for automated extra skeletal bone sarcoma classification," Materials Today: Proceedings from International Conference on Processing of Materials, Minerals and Energy (PMME 2016), vol. 5, pp. 1879-1889, 2018.

[30] Y. Li, M. Y. A. Khan, Y. Jiang, F. Tian, W. Liao, S. Fu et al., "CART and PSO + KNN algorithms to estimate the impact of water level change on water quality in Poyang Lake, China," Arabian Journal of Geosciences, vol. 12, no. 9, pp. 1-12, 2019.

[31] R. S. El-Sayed, "Linear discriminant analysis for an efficient diagnosis of heart disease via attribute filtering based on genetic algorithm," Journal of Computers, vol. 13, no. 11, pp. 1290-1299, 2018.

[32] S. Nagpal, S. Arora, S. Dey, and S. Shreya, "Feature selection using gravitational search algorithm for biomedical data," 7th International Conference on Advances in Computing & Communications, ICACC, vol. 115, pp. 258-265, 2017.

[33] C. Gunavathi and K. Premalatha, "Performance analysis of genetic algorithm with knn and svm for feature selection in tumor classification," International Journal of Computer and Information Engineering, vol. 9, no. 4, pp. 513-519, 2019.

[34] E. Sugiyarti, K. A. Jasmi, B. Basiron, M. Huda, S. K, and A. Maseleno, "Decision support system for scholarship grantee selection using data mining," International Journal of Pure and Applied Mathematics, vol. 119, no. 15, pp. 2239-2249, 2018.

[35] H. Rao et al., "Feature selection based on artificial bee colony and gradient boosting decision tree," Applied Soft Computing Journal, vol. 74, pp. 634-642, January 2019.

Hybrid Algorithm Naive Bayesian Classifier and Random Key Cuckoo Search for Virtual Machine Consolidation Problem

Yasser Moaly[1]
Department of Computer Science
Faculty of FCAI Cairo University, Cairo, Egypt

Basheer A.Youssef[2]
Doctor, Department of Computer Science
Faculty of FCAI Cairo University, Cairo, Egypt

***Abstract*—The trade-off between Energy consumption and SLA violation presents a serious challenge in cloud computing environments. A non-aggressive virtual machine consolidation algorithm is a good approach to reduce the consumed energy as well as SLA violation. A well-known strategy to deal with the virtual machine consolidation problem consists of four steps: host overloading detection, host under-loading detection, virtual machine selection and virtual machine placement. In this paper, the previous strategy is modified by merging the last two steps virtual machine selection and virtual machine placement, to avoid any poor solutions caused by solving both steps separately. In the host overloading/under-loading detection steps, we classified host status into five classes: Over-Utilized, Nearly Over-Utilized, Normal Utilized, Under-Utilized and Switched Off, then an algorithm, based on the Naive Bayesian Classifier, was introduced in order to detect the future host state for minimizing the number of virtual machine migrations; as a result, the energy consumption and performance degradation due to migrations will be minimized. In the virtual machine selection and placement steps, we introduced an algorithm based on the Random Key Cuckoo Search to reduce the energy consumption and enhance the SLA violation. To assess the algorithm, real data traces for 10 days, were used to verify the proposed algorithms. The experimental results proved that the proposed algorithms can significantly reduce the consumed energy as well as the SLA violation in data centers.**

Keywords*—Cloud computing; Naive Bayesian classifier; Random Key Cuckoo Search; Energy-efficiency; SLA-aware; Virtual Machine consolidation

I. Introduction

Pay-as-you-go basis [1] [2] has increased demand in Cloud computing (CC), especially after the important role played by big data and the internet of things (IOT). In addition, Cloud datacenters consume huge energy, which is mostly produced by non-green energy, resulting in a lot of carbon emissions as well as a high service cost. According to the United States Data Center Energy Usage Report in 2016, in 2010, the energy consumed by U.S data centers constituted about 2% of the global energy consumption [3]. Additionally, a six-month study that was conducted by Barroso & Hölzle (2007) showed that 5000 physical machines (PMs) used nearly around 10-50% of their total capacity [4].

A non-aggressive virtual machine (VM) consolidation is a good approach for reducing energy consumption with attention to the SLA violation (SLAV). VM consolidation is the technique of reallocating virtual machines (VMs) on PMs, subject to minimizing the used resources while keeping the other resources in the sleep mode. Nonetheless, one problem arises, i.e. the VM consolidation issue is an NP-Hard problem [5]. To solve the VM consolidation problem Beloglazov & Buyya (2012) proposed an effective strategy, consisting of four steps [6]: host overloading detection (HOD), host under-loading detection (HUD), VM selection (VMS) and VM placement (VMP). This strategy has been used by most researchers to handle the VM consolidation problem. However, this research found that the week point of this strategy is that it solves both VMS and VMP problems separately, which ultimately produces a poor solution. Therefore, in this study, the aforesaid strategy has been modified through merging the steps of VMS and VMP into a single step. The work in [7], [8] has shown that there is a linear relationship between CPU utilization and power consumption. Thus, in the HOD & HUD detection, it is proposed, using the PM State detection based on the Naive Bayesian classifier (PMSDNBC) [9], to detect the future state of the PM, based on the past CPU historical usage.

Within this study, the PM status is classified into five classes: Over-Utilized (OU), Nearly Over-Utilized (NOU), Normally Utilized (NU), Under-Utilized (UU) and Switched off (SO). The proposed prediction model will help in minimizing the VM migration, resulting in minimizing the energy consumption as well as performance degradation due to migrations. For VMS and VMP, a new algorithm called VM selection and placement using Random Key Cuckoo search (VMSPRKCS), was proposed to handle both VMS and VMP steps. Random Key Cuckoo Search (RKCS) [10] is a novel approach using Cuckoo Search (CS) [11], based on random keys encoding schema (RKES) [12]. CloudSim [13], with a real data set presented from PlanetLab [14], has been used to evaluate the proposed algorithm, and the experimental results showed that the proposed algorithm significantly reduces the consumed energy as well as SLAV.

The rest of the paper is organized as follows: the related work is to be discussed in Section II, a power consumption model in Section III, the Random Key Cuckoo Search concept in Section IV, the proposed overload detection as well as VM selection and placement algorithms in Section V, Simulation Setup and Results in Section VI.

II. Related Work

Since VM consolidation problem is regarded as an NP-Hard problem, it remains an interesting area for researchers, who need to work a lot more towards achieving optimization. Some approaches have used bin-packing optimization algorithms to handle the problem. Shi, Furlong and Wang [15] proposed a greedy algorithm based on the First-fit Decreasing (FFD) algorithm in attempts to reduce the energy consumed by datacenters. FFD has been widely used to find a near optimal solution to the vector pin backing problem. However, the proposed algorithm has a high performance, it can easily be stuck in a local optimum.

Keller, Tighe, Lutfiyya and Bauer [8], proposed new six placement policies based on combinations of order strategies for VMs and PMs. For PMs, they used 3 order strategies: (1) Increasing, which orders the partially-utilized and underutilized PMs in an increasing order according to the CPU utilization; (2) Decreasing, which orders the partially-utilized and underutilized PMs in a decreasing order according to the CPU utilization; and (3) Mixed, which orders the underutilized PMs in a decreasing order and the partially-utilized hosts in an increasing order according to the CPU utilization. For VMs, they only used two strategies: (1) the Decreasing strategy to order all VMs in a decreasing order according to the CPU load; and (2) the Increasing strategy to order all VMs in an increasing order according to the CPU load. The DCSim simulator [16], [17] has been used to evaluate the proposed policies, and the results showed that ordering PMs in a decreasing order and VMs in an increasing order according to the CPU utilization performs better in terms of the energy reduction than the other proposed policies.

Some other approaches have used Artificial Intelligence (AI) prediction models to predict the future PM state in attempts to reduce the energy consumption, the number of VM migrations and performance degradation due to migration. Lei Qiao,Bo Liu,Yang Hua,Qing Zhao & Xiong Fu [18], presented an algorithm based on the Genetic Expression Programming (GEP) [19] to consolidate VMs according to historical data. The algorithm focuses only on 2 steps, VM selection and VM placement. The algorithm carries out dynamic VM migrations for overloaded PMs and under loaded PMs, taking into consideration the proposed prediction model. The algorithm is compared to the most common algorithm Power- aware Best-fit Decreasing (PABFD) algorithm [6], and the results showed that the presented algorithm has a significant improvement in Service Level Agreement Time per Active Host (SLATAH) and in minimizing the number of virtual machine consolidations. Lianpeng Li,Jian Dong,Decheng Zuo & Jin Wu [20] built a prediction model based on Robust Simple Linear Regression (RobustSLR) to detect the future CPU utilization of hosts; then, they proposed overload/under load detection and VMP algorithms based on the proposed prediction model. In the VMP, they modified the PABFD algorithm in order to check the future host state using the proposed prediction model; in case, PM future state will be overloaded, and thus no VMs will be migrated to it. The proposed algorithm was assessed by the CloudSim toolkit with a real data presented from PlanetLab, and the results showed that the presented algorithm could reduce the energy consumption by at most 25.43% and SLA violations by at most 99.16%, compared with PABFD and the most common VM state detection & selection policies.

Lianpeng Li,Jian Dong,Decheng Zuo & JIaxi Liu [21] proposed a prediction model based on the Simple Bayesian Classifier to detect the future host state. They used the mean function to convert the CPU utilization history for the last hour, and then used them as the Bayes features and the host states (overloaded or not overloaded) as Bayesian Classifier labels. Then, they calculated the posterior probability using the prior probability and condition probability, and then the classifier predicted that that input vector belonged to the class having the highest posterior probability. The algorithm has been compared with the most common PDFA algorithm, and the results showed that the proposed algorithm performed better in terms of energy consumption and SLAV than did PDFA.

Other approaches have used meta-heuristic algorithms to find a near optimal solution. Dabiah Ahmed Alboaneen, Huaglory Tianfield & Yan Zhang [22], applied the Glowworm Swarm Optimization (GSO) Algorithm for the VMP problem. They used a CloudSim toolkit with real data presented from PlanLab to assess the algorithm. The results showed that GSO can perform better than many common placement policies. Khaoula BRAIKI & Habib YOUSSEF [23], proposed a multi-objective algorithm based on the Particle Swarm Optimization (PSO) algorithm to improve resource utilization while minimizing the energy consumption. Sanaz Tavakoli-Someh & Mohammad Hossein Rezvani [24], proposed a multi-objective NSGA-II meta heuristic algorithm to handle the objectives of: (1) maximizing the resources utilization; and (2) minimizing the number of used PMs. The experimental result showed that the proposed algorithm performed better than such basic approaches as the FFD algorithm and the best-fit decreasing (BFD) schemes.

Another statistical analysis approach, based on historical data, has been proposed by Beloglazov and Buyya [6]. It provides four overload host detection policies: Local Regression (LR), Median Absolute Deviation (MAD), Robust Local Regression (RLR) and Interquartile Range (IQR), in addition to 3 selection policies: Minimum Migration Time (MMT), Random Selection (RS) and Maximum Correlation (MC). The CloudSim simulator with PlanetLab traces has been used to evaluate the proposed overload state detection and VM selection policies. As shown by the results, the LR overload detection policy with MMT selection policy significantly surpasses the other policies in reducing SLAV and the number of VM migrations. That is why we considered LR_MMT 1.2 as the benchmark and compared the proposed approach in this research with it using the same traces provided from PlanetLab.

III. Power Consumption Model

According to different studies, the most energy consumed by servers is mostly consumed by the CPU rather than other resources [7], [25]. Hence, for simplicity, we can represent the PM energy consumption with its CPU utilization. We used a real power consumption data model, illustrated in Table I provided by SPEC power [26] for two servers, HP ProLiant ML110 G4 and HP ProLiant ML110 G5, to calculate the energy consumption.

TABLE. I. SPECPOWER POWER CONSUMPTION OF TWO SERVERS: HP PROLIANT G4 & G5 AT DIFFERENT LOAD LEVELS IN WATTS

CPU	**HP ProLiant G4**	**HP ProLiant G5**
0%	86	93.7
10%	89.4	97
20%	92.6	101
30%	96	105
40%	99.5	110
50%	102	116
60%	106	121
70%	108	125
80%	112	129
90%	114	133
100%	117	135

IV. THE RANDOM KEY CUCKOO SEARCH (RKCS) CONCEPT

The RKCS is a novel approach of Cuckoo Search Meta-heuristic algorithms based on a random key encoding schema. It is very effective in the combinatorial problems, as it will be used to apply Lévy flight on the old solution to generate a new cuckoo solution from it.

A. Cuckoo Search (CS)

The CS is a population based meta-heuristic optimization algorithm inspired by the natural behavior of cuckoo birds. It was presented by Xin-She Yang and Suash Deb in 2009. When cuckoo birds want to lay eggs, they do not build their own nests, but they rather use some other birds' nest. In case that other bird discovered that those eggs were not their own, it would either throw them or abandon its nest and build another one elsewhere. Below are the laws CS:

1) Number of hosts is fixed.

2) A nest can host only one egg.

3) Each parasitic egg represents a solution in the search space.

4) The transaction of generating a new solution (Cuckoo one) from old solution is accomplished using Lévy flight.

5) Height quality solutions with good fineness values will be carried forward to the next generation.

6) There is a probability pa [0, 1] to discover the parasitic foreign egg by the host bird.

7) For a cuckoo i, the new parasitic cuckoo (solutions) Xi(t+1) generated by using Lévy flight using Equation #1 where $\alpha > 0$ is the step size, and product $\oplus$ means entry-wise multiplication.

8) Lévy flight is a random step length in a random angle in which the random step length is generated from the Lévy distribution using equation #2.

$$x^{(t+1)} = x^{(t)} + \alpha \oplus \text{ Levy}(\lambda)\alpha \quad (1)$$

$$\text{Lévy} \sim u = x^{-\lambda}, (1 < \lambda < 3) \quad (2)$$

B. Random-Key Encoding Scheme (RKES)

The RKES is a transformation technique which is used to represent a vector in a continuous search space in a combinatorial form. The conception of RKES is to generate a random weight value from [0, 1] to each item in the array vector, and then order it in an ascending order according to the weight values. Table II illustrates a sample RKES.

TABLE. II. RANDOM-KEY ENCODING SCHEME (RKES)

WEIGHTS	0.8	0.6	0.7	0.3	0.1	0.4
DECODED AS	6	3	4	2	4	7

V. PROPOSED OVERLOAD / UNDERLOAD DETECTION, VM SELECTION AND VM PLACEMENT ALGORITHMS

An effective strategy proposed by Beloglazov and Buyya (2012) [6] to deal with the VM consolidation problem consists of the following four steps:

1) Host Overloading Detection (HOD): to detect the overloaded PMs; some VMs must be migrated from them to eliminate SLAV.

2) Host Under-loading Detection (HUD): to detect PMs with low utilization resources; all VMs must be migrated from them if possible, and switch them into the sleep mode.

3) Virtual Machine Selection (VMS): responsible for selecting some VMs from overloaded PMs in order to migrate them to another suitable PMs.

4) Virtual Machine Placement (VMP): responsible for finding a new suitable PMs for the selected VMs.

In this research, we merged the steps, VMS and VMP into one to avoid any poor solution resulting from solving both separately. In the HOD & HUD steps, a prediction algorithm based on the Naive Bayesian Classifier is proposed to predict the future host state for minimizing the VMs migrations. In the VMS and VMP steps, we proposed an algorithm based on the RKCS to reduce the energy consumption and SLAV. Solution construction, host state prediction model, and the proposed VM selection as well as placement policy are discussed in the following sections.

A. Solution Construction

Assume that we have n VMs and m PMs; in the context of VM consolidation problem and RKCS, a single egg solution S_i is represented by: the VMs random Keys ($VMRK_i$), the PMs random keys ($PMRK_i$) and the VM migrations mappings ($VMMM_i$), which are constructed as shown in Tables III, IV and V, respectively.

Each element in $VMRK_i$ will have a VM index (VMI_i) and VM weight value (VMW_i). Each item in $PMRK_i$ will have: PM Index (PMI_i), PM Status weigh value ($PMSW_i$) and PM weight value (PMW_i). Finally, $VMMM_i$ represents the placement mappings; each value in the array represents a PM index to which will the VM will be placed. For instance, if the second element of array is 7, then the second VM will be migrated to PM of index 7.

TABLE. III. VIRTUAL MACHINES RANDOM KEY

VMI_1	VMI_2		VMI_3	VMI_4	…	VMI_n
VMW_1	VMW_2		VMW_3	VMW_4	…	VMW_n

TABLE. IV. PHYSIAL MACHINES RANDOM KEYS

PMI_1	PMI_2	PMI_3	PMI_4	…	PMI_m
$PMSW_1$	$PMSW_2$	$PMSW_3$	$PMSW_4$	…	$PMSW_m$
PMW_1	PMW_2	PMW_3	PMW_4	…	PMW_m

TABLE. V. VIRTUAL MACHINES MIGRATION MAPPINGS

VM_1	VM_2	VM_3	VM_4	…	VM_n
PMI_1	PMI_2	PMI_3	PMI_4	…	PMI_n

B. Physical Machine State Prediction Model using Naive Bayesian Classifier (PMSDNBC)

The study presented by Haikun Liu, Hai Jin, Cheng-Zhong Xu & Xiaofei Liao [27], stated that live migration is not free, and that is why predicting the PM future state will play a significant role in minimizing the number of VM migrations, resulting in minimizing energy consumption and performance degradation due to the migration. In this research, a new algorithm called PM state predection using Naive Bayesian classifier (PMSDNBC) was introduced to detect the future host state. The Naive Bayesian classifier is an AI machine learning model used to classify different objects based on certain features. In the current research, the model is used to predict the PM future state according to the past CPU utilization history. For the classifier features, we constructed n+1 dimensional feature vector using the latest CPU utilization history CPU_t, CPU_{t-1}, CPU_{t-2}, CPU_{t-3}, CPU_{t-4}.. CPU_{t-n+1} in the time t, t-1, t-2, t-3, t-4 … t-n+1 for the last n preceding points. Then, we got input vector F = [CPU_t, CPU_{t-1}, CPU_{t-2}, CPU_{t-3}, CPU_{t-4}... , CPU_{t-n+1}], and basically, as we need to detect the future state of PM, we used the PM states as the Bayes classifier labels. In most researches the PM status is classified into the following three states: (1) Over-Utilized, (2) Under-Utilized and (3) Switched Off. For the Over-Utilized PMs, some VMs must be migrated from them to other non Over-Utilized PMs in order not to elimnate the SLA. For the Under-Utilized PMs, all PMs must be migrated from them, if possible, in order to switch them to the sleep mode. In this reserach we found the below weak points in this classification:

- PMs which will accept migrations from the Over-Utilized PMs and PMs from which all VMs must be migrated, if possible, are both treated as "Under-Utilized".
- In the previous classification, there is no state to describe the PMs which are not Over-Utilized and are not able to accept new migrations, as they are near to be Over-Utilized.

For this reason, in this research we classified the PM statuses into the below five classes:

1) *Over-Utilized (OU):* PMs which are over utilized, so some VMs must be migrated from them in order not to violate SLA.

2) *Nearly Over-Utilized (NOU):* PMs which are near to be over-utilized. These types of PM will not accept any migrations nor migrate VMs from them.

3) *Normal Utilized (NU):* PMs which fall between nearly over-utilized and underutilized thresholds. These types of PMs will accept new migrations but not VMS will be migrated from them.

4) *Under-Utilized (UU):* PMs which are underutilized; all VMs hosted on these PMs should be migrated away to other suitable PMs if possible, then switch these UU PMs to the sleep mode.

5) *Switched Off (SO):* Switched off PMs.

Accordingly, classifier labels will be L = {OU, NOU, NU, UU, SO}. Static threshold values based on the experimental results have been used to identify: OU, NOU, NU, UU and SO thresholds; the values are illustrated in Table VI. Classifier dataset is constructed according to the below rules:

1) *In case the CPU* utilization value equals 0, the value of “SO” label will be set to 1 and the other labels will be set to minimum product effect value (MPEV).

2) *In case the CPU* utilization value is greater than the over-utilized threshold, the value of the “OU” label will be set to 1 and the other labels will be set to MPEV.

3) *Otherwise*, we will calculate the percentage of how far the current state from the upper and lower thresholds using equations #3 & #4 and the reset labels are set to MPEV.

$$\mathrm{VUTi} = \frac{(\mathrm{UPT}-\mathrm{LWT})-(\mathrm{UPT}-\mathrm{CPUi})}{(\mathrm{UPT}-\mathrm{LWT})} \tag{3}$$

$$\mathrm{VLTi} = \frac{(\mathrm{UPT}-\mathrm{LWT})-(\mathrm{CPUi}-\mathrm{LWT})}{(\mathrm{UPT}-\mathrm{LWT})} \tag{4}$$

Where UPT is the upper threshold value, LWT is the lower threshold value, VUTi is the value of state i for the upper label and VLTi is the value of state i for the lower label.

Algorithm 1: PMSDNBC

Input: CPU Utilization History
Output: Host State

```
n = 10; minValue=0.0001;
For i =0 to n − 1 do
    F[i] = cpuUtilizationHistory [n − i − 1];
    IF F[i] == 0 Then
        Set switchedOffValues[i] =1;
        Set Other Values to minValue;
    Else If F[i] > overUtilizedThreadShold Then
        Set overUtilizedValues [i] =1;
        Set Other Values to minValue;
    Else
        Calculate VUTi value equation #1;
        Calculate VLTi value equation #2;
        Set other labels values to minValue;
    End If;
Calculate P(x | OU), P(x | NOU), P(x | NU), P(x | UU)
    & P(x | SO) using equation #5.
Return state of highest probability
```

TABLE. VI. STATIC THRESHOLDS VALUES

THRESHOLD	RANGE
OVER-UTILIZED	CPU >= 0.8
NEARLY OVER-UTILIZED	0.8 > CPU >=0.78
NORMAL UTILIZED	0.78> CPU >=0.5
UNDER-UTILIZED	0.5> CPU > 0
SWITCHED OFF	CPU = 0

The value 0.00001 has been used as MPE, first, to avoid zero values generated from product operation and second, because it has lower effect in the product operation.

Finally, for each label l in L, the probability of P (x|l) is calculated using equation #5 and according to the Naive Bayesian Classifier, PM state will be classified as the state with the highest probability.

$$p(x|l) = \prod_{i=1}^{n} P(fi|l) \quad (5)$$

$$P(fi|l) = \frac{P(fi \cap l)}{P(l)} \quad (6)$$

C. Virtual Machine Selection and Placement using Random Key Cuckoo Search (VMSPRKCS)

VMS & VMP is the process of selecting some VMs from over-utilized and under-utilized PMs and finding a new placement for them. In this research, an algorithm based on RKCS is proposed to handle the two steps VMS & VMP into one step to avoid producing a poor solution resulting from solving both steps separately. First, we detect the state of all PMs using the proposed algorithm PMSDNBC. Then, we added all VMs of over-utilized and under-utilized PMs into one dimensional vector VMs = [V1, V2, V3 … Vn]. For the over-utilized PMs, we will migrate some VMs from them in order to not to violate SLA; for the under-utilized PMs, all VMs will be migrated from them if possible in order to switch them to the sleep mode. Since RKCS is a population based algorithm, we will have n solutions S = [s_1, s_2, s_3 … s_n], and for each solution s_i, we will have $VMRK_i$ and $PMRK_i$. The idea behind VMSPRKCS is consists of the below steps:

1) Detect the status of all PMs using PMSDNBC algorithm.

2) Add all VMs of over-utilized & under-utilized PMs into one dimensional array SVMs = [VM1,VM2,….VMn].

3) Assigning a weight value for each VM in SVMs, either generated randomly between [0-1] or by Lévy using equation #1 when generating a cuckoo egg from an old one.

4) Assigning a weight value and a status weight value for each PM. The PM weight value is either generated randomly between [0-1] or by Lévy using equation #1 from an old solution. The PM status weight value is generated according to Table VII.

5) Sorting VMs in SVMs array in an ascending order by the weight value

6) Sorting PMs in an ascending order by the status weight value then by the weight value.

7) Each VM will be allocated to the first PM fitting for it, which will not be overloaded after the migration.

8) In case the status of the old PM, from which the VM is migrated, changed from "Over-Utilized" to any new status, then the remaining VMs on the old PM added to an excluded list in order to skip their migration.

9) In case the status of the new PM, to which the VM is migrated, changed to any new status, then new PM status weight value is recalculated based on PMSDNBC algorithm and Table VII, then all PMs are reordered in an ascending order by their status weight value then by the weight value.

TABLE. VII. PM STATUS WEIGHT INDEX MAPPINGS

STATUS	STATUS WEIGHT VALUE
NORMAL UTILIZED	1
UNDER-UTILIZED	2
SWITCHED OFF	3
NEARLY OVER-UTILIZED	4
OVER-UTILIZED	5

Algorithm 2: VMSPRKCS

Input: PmList, VmList
Output: VMs migration mappings

```
For i=0 to PmList.length Do
    Let pmStatus = getHostStatus(pmList [i])
    If pmStatus=='Over-Utilized' or pmStatus=='Under-Utilized' then
        vms.add(pmList [i].getVmList());
    End If
End for
return FindBestAllocation(PmList, vms);
```

Algorithm 3: FindBestAllocation

Input: PMs, VMs
Output: VMs migration mapping

```
n=10;
Generate an initial population of n host nests;
While (iterations < maxIteration) do
    Get a cuckoo randomly solution Si by L´evy flights
    ApplyPlacement (pms, vms, Si)
    Calculate Fi quality of Si
    Choose a nest among n (say, j) randomly
    If (Fi > Fj ) then
        replace j by the new solution;
    end
A fraction (pa) of worse nests is abandoned and new ones are built.
Keep the best solutions
Rank the solutions and find the current best
End while
Return best;
```

Algorithm 4: ApplyPlacement

Input: PMs, VMs, S

```
Sort S.VMRK in ascending order by Weight Value
Sort S.PMRK in ascending order according by Status Weight value then by Weight Value
Let VMRK = S.VMRK
For i=0 to VMRK .length do
    Let vm = VMs[VMRK[i].index];
    If ExcludedVMs .contains(vm) then
        Continue;
    End If
    Let pm = vm.getHost();
    Let pmPreStatus = PMSDNBC(pm);
    pm.vmDestroy(vm)
    Let pmPostStatus = PMSDNBC(pm);
    IF pmPreStatus == "Over Utilized" and pmPostStatus!=  "OverUtilized" then
        ExecludedVms.addAll(pm.getVmList());
    End If
    FindFirstFittingPm(PMs, vm, S)
End For
End For
```

Algorithm 5: FindFirstFittingPm

```
Input: PMs, VM, S
Output: PM
Let selectedPm = null;
Let PMRK = S. PMRK;
For j=0 to PMRK .length do
    Let pm = pms[PMRK[i].index];
    If isHostOverUtilizedAfterAllocation(pm,VM) then
        continue;
    Else
        Let pmPreStatus = PMSDNBC(pm);
        pm.vmCreate(vm);
        Let pmPostStatus = PMSDNBC (pm);
        If pmPreStatus!= pmPostStatus then
            Get pm weight status value using PMSDNBC
            and Table VII;
            Sort PMRK in ascending order according by
            Status Weight then by Weight Value;
        End If
        selectedPm = PM;
        break;
    End If
End For
Return selectedPm ;
```

VI. Simulation Setup and Results

Performance metrics, Simulation setup & Experimental results are discussed in the following sections.

A. Performance Metrics

The following metrics have been considered to evaluate the proposed algorithm:

1) Energy consumption: it refers to the total energy consumed by all PMs in the datacenter, where the PM energy consumption is calculated according to a real data power consumption model provided by the SPEC power benchmark. Table I illustrates the energy consumption of two different PMs, HP G4 and HP G5, at different CPU load levels.

2) Service Level Agreement Violation (SLAV): As proposed in [6], SLAV can be measured in IaaS for two main factors: (1) SLA violation time per active host (SLATAH) resulting from the CPU utilization is of 100% and (2) the SLA violations resulting from performance degradation due to migration (PDM). The SLAV factors can be calculated using equations #7.

$$\mathrm{SLAV} = \mathrm{SLATAH} * \mathrm{PDM} \tag{7}$$

$$\mathrm{SLATAH} = \frac{1}{\mathrm{m}}\sum_{\mathrm{k}=0}^{\mathrm{m}} \frac{\mathrm{Tok}}{\mathrm{Tak}} \tag{8}$$

Where m is the number of PMs, Tok is the total time in which the PM had an over-utilized status, resulting from the 100% CPU utilization and Tak is the total time of the PM for being in the active state.

$$\mathrm{PDM} = \frac{1}{\mathrm{n}}\sum_{\mathrm{k}=0}^{\mathrm{n}} \frac{\mathrm{Cdk}}{\mathrm{Crk}} \tag{9}$$

Where n is the number of VMs, Cdk is the estimated performance degradation of the VM k due to migrations and Crk is the total requested CPU capacity by the VM j.

1) ESV: as the research target is to balance the trade-off between the SLA violation and energy consumption, it is important to consider this metric. It is simply calculated as the product of SLAV and energy consumption

$$ESV = SLAV * Enegry\_consumption \tag{10}$$

2) Number of VM migrations: it refers to the total number of VM migrations occurred over all the datacenter.

3) Shutdown hosts: it refers to the total number of PMs switched to the switched off mode over all the datacenter.

B. Simulation Setup

To evaluate the proposed algorithm, we used the CloudSim 3.0.3 toolkit simulator. Cloudsim is a well-known common simulator that supports different policies for host overload detection, VM selection and VM placement. It also provides different types of workload as well as several cloud metrics calculation, such as: Energy Consumption, SLAV, number of VM migrations, PDMA, SLATAH and number of host shutdowns. Furthermore, we used real workload traces from a real system (PlanetLab data). PlanetLab is the monitoring part of the CoMon project. It monitored CPU utilizations for more than thousand VMs hosted at more than 500 PMs which were collected during March and April 2011. Each day in the traces has a file for each VM, containing 288 values which represent the VM CPU utilization value [0-100] every 5 minutes during the day. Traces characteristics are represented in Table VIII. A datacenter comprising 800 heterogeneous PMs and more than 1000 VMs was simulated; half of the PMs were HP ProLiant ML110 G4 (Intel Xeon 3040, dual-core 1860 MHz, 4 GB, 1 Gbps) and the rest are HP ProLiant ML110 G5 (Intel Xeon 3075, dual-core 2660, 4 GB, 1 Gbps). For the VMs, four types were used, corresponding to Amazon EC2 [27] illustrated below:

1) Micro instance (613MB, 500 MIPS).
2) Small Instance (1.7 GB, 1000 MIPS).
3) Extra-large Instance (3.75 GB, 2500 MIPS).
4) High-CPU Medium Instance (0.85 GB, 2500 MIPS).

C. Experimental Results

Traces with heterogeneous states for real cloud datacenter presented from PlanetLab illustrated in Table VIII have been used to evaluate the algorithm. A study presented by Beloglazov and Buyya [6], stated that lr_MMT 1.2 performs better than other dynamic VM consolidation algorithms, so we considered it as the benchmark and compared our proposed algorithm with it. The results are illustrated in Fig. 1, 2, 3, 4, 5, 6 and 7. The experimental results show that the proposed algorithm can highly reduce the below metrics:

1) Energy consumption reduced by minimum 17.7%, by maximum 28.6% and with average 24.23%, compared with lr_MMT 1.2.

2) SLAV reduced by minimum 93.52%, by maximum 96.89% and with average 95.35%, compared with lr_MMT 1.2.

3) ESV reduced by minimum 95.5%, by maximum 97.56% and with average 96.5%, compared with lr_MMT 1.2.

4) The number of VM migrations reduced by minimum 88.07%, by maximum 90.85% and with average 89.3%, compared with lr_MMT 1.2.

5) PDM reduced by minimum 77.78%, by maximum 88.89% and with average 86.3%, compared with lr_MMT 1.2.

6) SLATAH reduced by minimum 60.08%, by maximum 81.9% and with average 71.07%, compared with lr_MMT 1.2.

7) Number of host shutdowns reduced by minimum 80.27%, by maximum 88.31% and with average 84.38%, compared with lr_MMT 1.2.

Finally, we ran the experiment for 10 times and calculated the median value, displaying it in terms of each performance metrics.

TABLE. VIII. PLANETLAB WORKLOAD TRACES CHARACTERISTICS

Date	No of Virtual Machines	Mean-Load (%)
03/03	1052	12.31
06/03	898	11.44
09/03	1061	10.70
22/03	1561	9.26
25/03	1078	10.56
03/04	1463	12.39
09/04	1358	11.12
11/04	1233	11.56
12/04	1054	11.54
20/04	1033	10.43

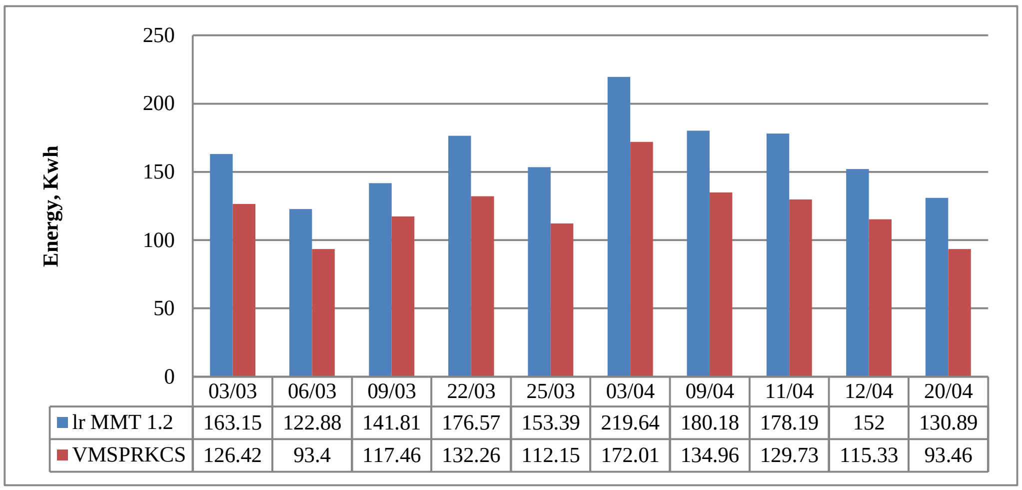


Fig. 1. Comparison of Energy Consumption.

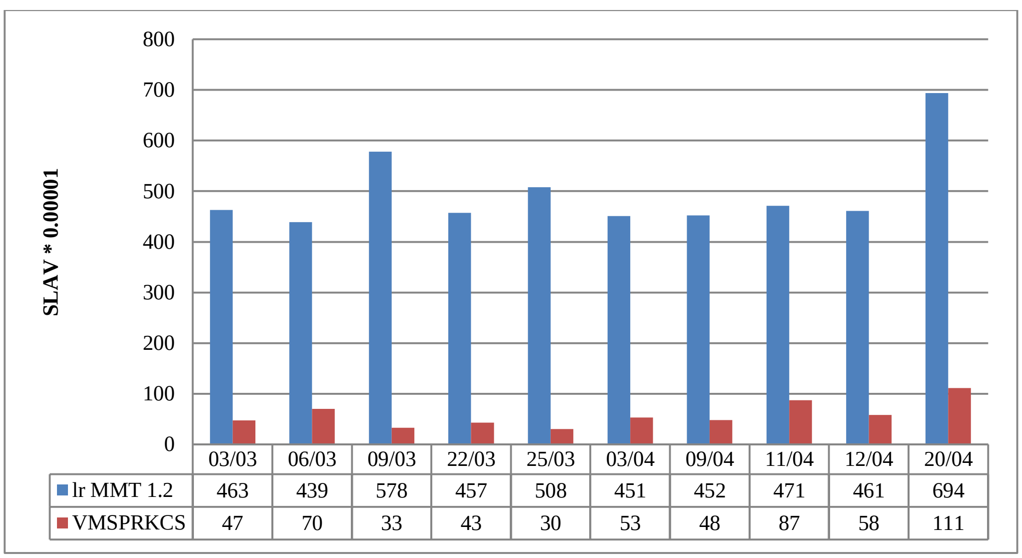
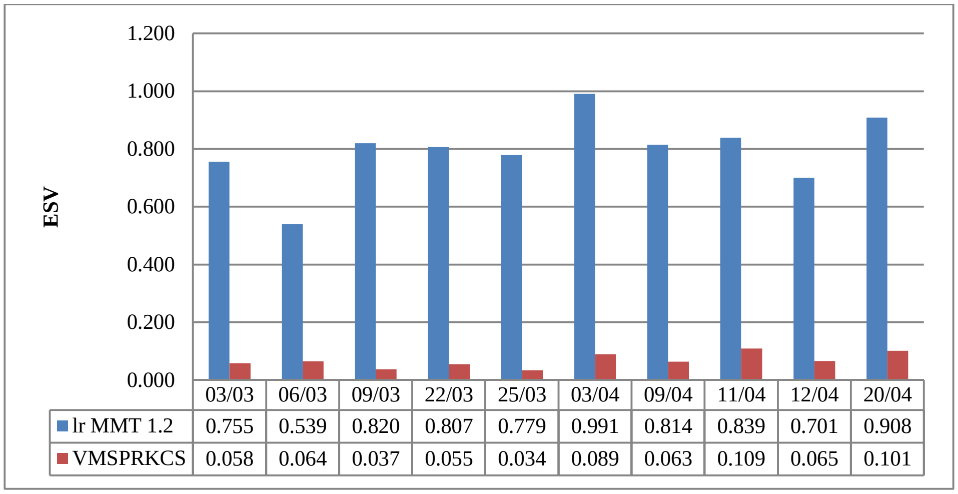
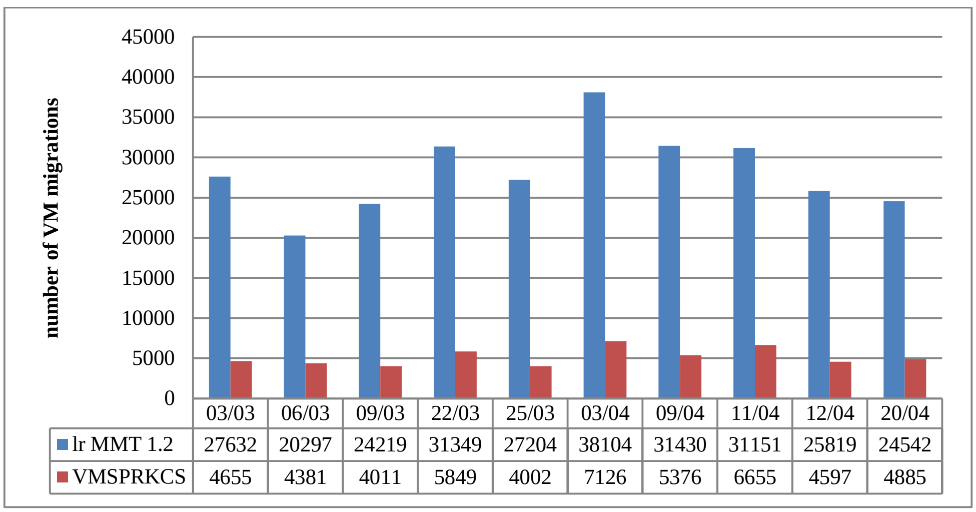


Fig. 2. Comparison of SLAV*0.0001.



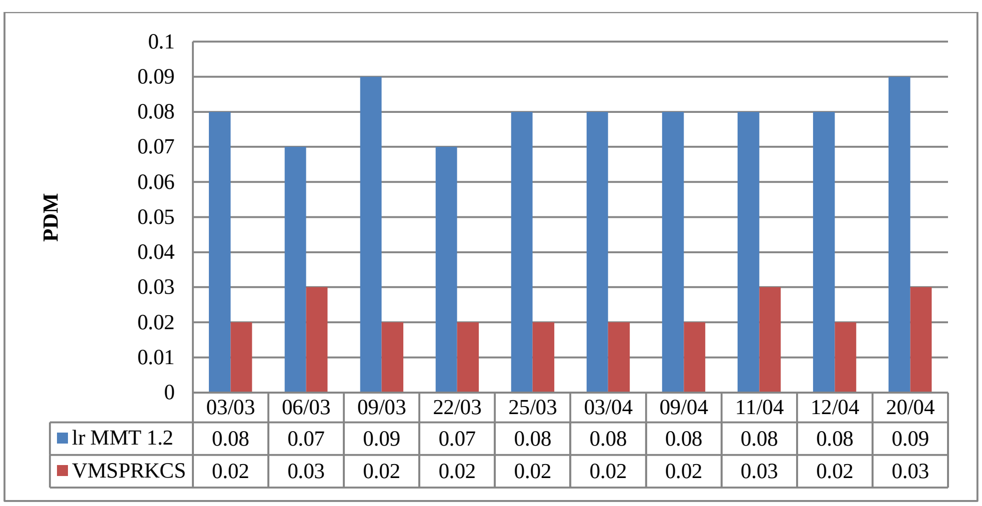
	03/03	06/03	09/03	22/03	25/03	03/04	09/04	11/04	12/04	20/04
lr MMT 1.2	0.755	0.539	0.820	0.807	0.779	0.991	0.814	0.839	0.701	0.908
VMSPRKCS	0.058	0.064	0.037	0.055	0.034	0.089	0.063	0.109	0.065	0.101

Fig. 3. Comparison ESV.



	03/03	06/03	09/03	22/03	25/03	03/04	09/04	11/04	12/04	20/04
lr MMT 1.2	27632	20297	24219	31349	27204	38104	31430	31151	25819	24542
VMSPRKCS	4655	4381	4011	5849	4002	7126	5376	6655	4597	4885

Fig. 4. Comparison of Number of VM Migrations.



	03/03	06/03	09/03	22/03	25/03	03/04	09/04	11/04	12/04	20/04
lr MMT 1.2	0.08	0.07	0.09	0.07	0.08	0.08	0.08	0.08	0.08	0.09
VMSPRKCS	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.03

Fig. 5. Comparison of PDM.

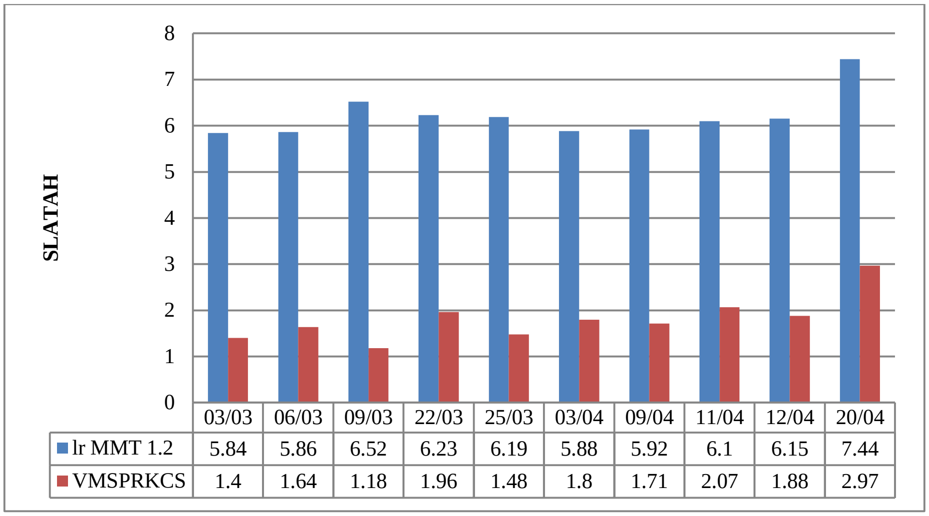


Fig. 6. Comparison of SLATAH.

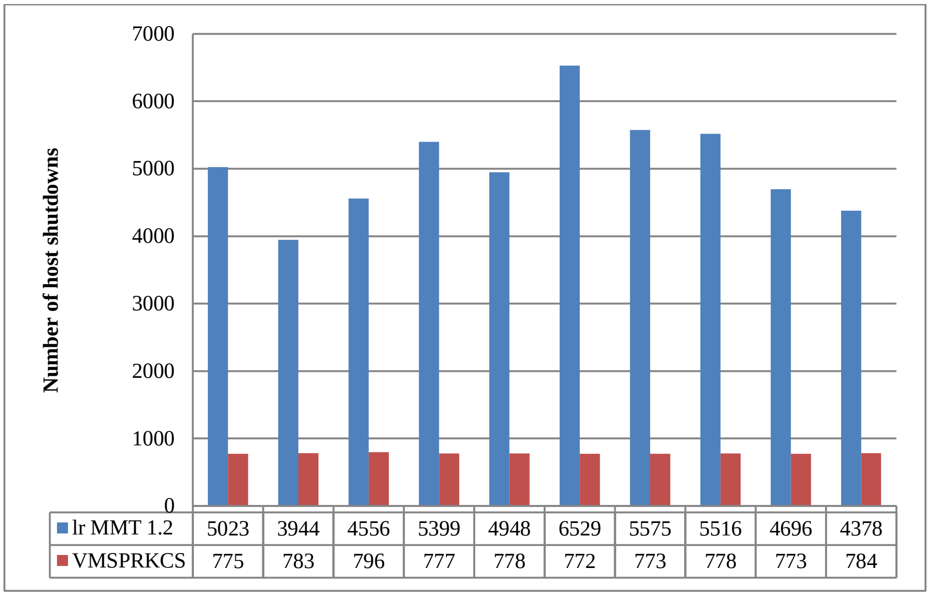


Fig. 7. Comparison of Number of Host Shutdowns.

VII. Conclusion and Future Work

In this paper, a hybrid algorithm based on the Naive Bayesian Classifier and the Random Key Cuckoo Search is introduced to balance the tradeoff between the energy consumption and SLA violation. In addition, we modified the most common strategy for handling the VM consolidation by merging VM selection and placement steps into one to avoid any poor solution that may arise due to solving each of the two steps alone. We used Naive Bayesian Classifier to detect the future PM state in order to minimize the VMs migration, resulting in reducing energy, SLAV as well as performance degradation due to migration. We used Random Key Cuckoo Search to handle the VM selection and placement steps. In addition, CloudSim has been used with real traces provided from PlanetLab to evaluate the proposed algorithm compared with the benchmark algorithm lr_MMT 1.2 and the results have shown that the proposed algorithm can reduce the energy consumption by 24.23%, SLAV by 95.35%, ESV by 96.5%, the number of VM migrations by 89.3%, PDM by 86.3%, SLATAH by 71.07% and the number of host shutdowns by 84.38%. In this research, the objective function was considered based data center energy consumption only; in the future we are interested in calculating the objective function according to multiple objective metrics and comparing it with other meta-heuristic algorithms as well as more real datasets.

References

[1] Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems, 25(6), 599–616. doi: 10.1016/j.future.2008.12.001.

[2] Mell, P. M., & Grance, T. (2011). The NIST definition of cloud computing. doi: 10.6028/nist.sp.800-145.

[3] Shehabi, A., Smith, S., Sartor, D., Brown, R., Herrlin, M., Koomey, J, Lintner, W. (2016). United States Data Center Energy Usage Report. doi: 10.2172/1372902.

[4] Barroso, L. A., & Hölzle, U. (2007). The Case for Energy-Proportional Computing. Computer, 40(12), 33–37. doi: 10.1109/mc.2007.443.

[5] Sharma, N. K., & Reddy, G. R. M. (2015). Novel energy efficient virtual machine allocation at data center using Genetic algorithm. 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN). doi: 10.1109/icscn.2015.7219897.

[6] Beloglazov, A., & Buyya, R. (2011). Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers. Concurrency and Computation: Practice and Experience, 24(13), 1397–1420. doi: 10.1002/cpe.1867.

[7] Fan, X., Weber, W.-D., & Barroso, L. A. (2007). Power provisioning for a warehouse-sized computer. Proceedings of the 34th Annual International Symposium on Computer Architecture - ISCA 07. doi: 10.1145/1250662.1250665.

[8] Keller, G., Tighe, M., Lutfiyya, H., & Bauer, M. (n.d.). An analysis of first fit heuristics for the virtual machine relocation problem. 2012 8th International Conference on Network and Service Management (Cnsm) and 2012 Workshop on Systems Virtualiztion Management (Svm), 406–413.

[9] Duda, R. O., Stork, D. G., & Hart, P. E. (2000). Pattern classification and scene analysis. New York: Wiley.

[10] Ouaarab, A., Ahiod, B., & Yang, X.-S. (2014). Random-key cuckoo search for the travelling salesman problem. Soft Computing, 19(4), 1099–1106. doi: 10.1007/s00500-014-1322-9.

[11] Yang, X.-S., & Deb, S. (2009). Cuckoo Search via Lévy flights. 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC). doi: 10.1109/nabic.2009.5393690.

[12] Bean, J. C. (1994). Genetic Algorithms and Random Keys for Sequencing and Optimization. ORSA Journal on Computing, 6(2), 154–160. doi: 10.1287/ijoc.6.2.154.

[13] Calheiros, R. N., Ranjan, R., Beloglazov, A., Rose, C. A. F. D., & Buyya, R. (2010). CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Software: Practice and Experience, 41(1), 23–50. doi: 10.1002/spe.995.

[14] Park, K., & Pai, V. S. (2006). CoMon. ACM SIGOPS Operating Systems Review, 40(1), 65. doi: 10.1145/1113361.1113374.

[15] Shi, L., Furlong, J., & Wang, R. (2013). Empirical evaluation of vector bin packing algorithms for energy efficient data centers. 2013 IEEE Symposium on Computers and Communications (ISCC). doi: 10.1109/iscc.2013.6754915.

[16] Tighe, M., Keller, G., Bauer, M., & Lutfiyya, H. (n.d.). DCSim: A data centre simulation tool for evaluating dynamic virtualized resource management. 2012 8th International Conference on Network and Service Management (Cnsm) and 2012 Workshop on Systems Virtualiztion Management (Svm).

[17] Digs-Uwo. (2014, December 7). digs-uwo/dcsim. Retrieved from https://github.com/digs-uwo/dcsim.

[18] Qiao, L., Liu, B., Hua, Y., Zhao, Q., & Fu, X. (2019). Genetic Expression Programming Based Dynamic Virtual Machine Consolidation in Cloud Computing. 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC). doi: 10.1109/iceiec.2019.8784588.

[19] Rui, Z., Shasha, H., & Hui, G. (2016). Research and Application of GEP Algorithm Based on Cloud Model. International Journal of Signal Processing, Image Processing and Pattern Recognition, 9(11), 309–318. doi: 10.14257/ijsip.2016.9.11.28.

[20] Li, L., Dong, J., Zuo, D., & Wu, J. (2019). SLA-Aware and Energy-Efficient VM Consolidation in Cloud Data Centers Using Robust Linear Regression Prediction Model. IEEE Access, 7, 9490–9500. doi: 10.1109/access.2019.2891567.

[21] Li, L., Dong, J., Zuo, D., & Liu, J. (2018). SLA-Aware and Energy-Efficient VM Consolidation in Cloud Data Centers Using Host States Naive Bayesian Prediction Model. 2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom). doi: 10.1109/bdcloud. 2018.00025.

[22] Alboaneen, D. A., Tianfield, H., & Zhang, Y. (2016). Glowworm Swarm Optimisation Algorithm for Virtual Machine Placement in Cloud Computing. 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld). doi: 10.1109/uic-atc-scalcom-cbdcom-iop-smartworld.2016.0129.

[23] Braiki, K., & Youssef, H. (2018). Multi-Objective Virtual Machine Placement Algorithm Based on Particle Swarm Optimization. 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC). doi: 10.1109/iwcmc.2018.8450527.

[24] Tavakoli-Someh, S., & Rezvani, M. H. (2019). Utilization-aware Virtual Network Function Placement Using NSGA-II Evolutionary Computing. 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI). doi: 10.1109/kbei.2019.8734978.

[25] Kusic, D., Kephart, J. O., Hanson, J. E., Kandasamy, N., & Jiang, G. (2008). Power and performance management of virtualized computing environments via lookahead control. Cluster Computing, 12(1), 1–15. doi: 10.1007/s10586-008-0070-y.

[26] The SPECpower benchmark. (n.d.). Retrieved from http://www.spec.org/power_ssj2008/.

[27] Liu, H., Xu, C.-Z., Jin, H., Gong, J., & Liao, X. (2011). Performance and energy modeling for live migration of virtual machines. Proceedings of the 20th International Symposium on High Performance Distributed Computing - HPDC 11. doi: 10.1145/1996130.1996154.

Adjacency Effects of Layered Clouds by Means of Monte Carlo Ray Tracing

Kohei Arai
Faculty of Science and Engineering
Saga University, Saga City, Japan

***Abstract*—Adjacency effects from layered box shaped clouds are clarified by means of Monte Carlo Simulation: MCS taking into account a phase function of cloud particles and multi-layered plane parallel atmosphere. MCS allows estimation of top of the atmosphere radiance. Influences on adjacency effects of phase function of the clouds in concern and the number of layers of the plane parallel atmosphere are also clarified together with the effects from the top and the bottom clouds. There are 10 of cloud types in meteorological definition. One-layer cloud for cumulus and cumulonimbus clouds are investigated in this study.**

Keywords—Monte Carlo simulation; top of the atmosphere radiance; cloud type; adjacency effects; layered clouds

I. Introduction

As described in [1], [2] the adjacent effect is a phenomenon in which reflected light components from other than the observation target are superimposed on the luminance from the target. For example, the brightness of an object with a small emission brightness, such as the sea surface, may be affected by several times the effect of cloud adjacency. Correction of this effect is important. According to Wilson A [3], when Turbid Case 2 seawater1 is targeted, the sea surface emission brightness in the visible wavelength range is 2-3.5 [μW / cm / str / nm], which is about 100 times that of high-rise clouds. It also extends. Therefore, the 1% adjacent effect is comparable to the sea surface emission brightness and cannot be ignored.

There are reports about the cloud adjacent effect by Monte Carlo method[2] already [4]. These are Monte Carlo simulations analysis of radiative transfer analytically assuming the ground surface, sea level, clouds, etc. in various atmospheric conditions

This Monte Carlo method was applied when it was impossible. In past studies, the effect of multiple reflections between the cloud base and the ground surface, the effect of scattering in clouds with different optical thicknesses, and the effect of multiple reflections between the cloud side and the ground surface were clarified by the Monte Carlo method. Clouds are layered. The effect of each layer on the peripheral effect of clouds is also clarified [5]. However, in these simulations, in order to shorten the time required for the Monte Carlo simulation, the atmosphere and clouds are represented by a one-layer model, taking into account the vertical nonuniformity of the real atmosphere and clouds.

The level of cloud droplets replacing the atmosphere with a multilayer parallel plate model [6].

The evaluation result of the cloud adjacency effect by Monte Carlo simulation using the same correlation number as the aerosol phase function has already been reported [7]. However, cloud distribution and cloud particle phase function were not considered. The phase function of cloud particles is the same as that of aerosol [8]. On the other hand, adjacency effect of layered clouds estimated with Monte-Carlo simulation is proposed and validated [9].

Evaluation of cirrus cloud detection accuracy of GOSAT/CAI and Landsat-8 with Laser Radar: Lidar is proposed and is validated with Calipso data [10]. In the meantime, comparative study on cloud parameter estimation among GOSAT/CAI, MODIS, CALIPSO/CALIOP and Landsat-8/OLI has been done with laser radar as truth data [11]. Meanwhile, thresholding based method for rain, cloud detection with NOAA/AVHRR data by means of Jacobi iteration method is proposed [12].

Typically, Radiative Transfer Equation: RTE has to be solved for the situations of which targeted scene includes clouds. It, however, is difficult to formulate RTE for such situation of remote sensing satellite data in particular for layered clouds. In this paper, a method for solving RTE for such situation (Layered clouds are included in the targeted scene of remote sensing satellite images by means of Monte Carlo Ray Tracing: MCRT.

The following section describes layered clouds in concern followed by Monte Carlo simulation (MCRT) in the presence of clouds in a uniform atmosphere. Then some results from simulation study is described followed by conclusion with some discussions.

II. Layered Clouds in Concern

When the cloud base height and the cloud top height are Ht and Hb, respectively, the 10 cloud types [9] defined by the World Meteorological Organization are as shown in Table I. In this study,

- One-layer cloud for cumulus and cumulonimbus clouds

1 https://www.researchgate.net/publication/232712020_Optical_Modeling_of_Ocean_Waters_Is_the_Case_1_Case_2_Classification_Still_Useful.

2 https://virial.com/monte-carlo-ray-tracing.html.

- Adjacent effects are evaluated for two-layered clouds, which are layered clouds such as lower-layer turbulent clouds, stratocumulus clouds, and middle-layer high-cumulus clouds and high-rise clouds. Fig. 1 schematically shows the one layer and two layer clouds assumed here.

A. One Layer Clouds

Cumulus clouds and cumulonimbus clouds, which are the targets of one layer of clouds, are clouds that develop vertically and are sometimes called convective clouds. Cumulus clouds can spread from near the ground to the middle and upper layers, and are often seen in the daytime when the weather is nice. The cumulonimbus is a dense cloud that grows vertically, and lightning and heavy rain fall from this cloud. Such lower clouds are considered to have a strong reflection effect that reduces sunlight by reflecting sunlight and reaching the ground surface. According to GLAS model [13], the albedo of cumulonimbus (Cb) is 0.8 and the albedo of cumulus (Cu) is 0.7. Therefore, this time, from the GLAS model of [13], the reflectance (R) of the cloud of one layer is 0.7 and the optical thickness (OD) is 9.0.

B. Two Layer Clouds

The high cumulus clouds that are the target of the upper part of the two-layer clouds are white or gray, rounded or rolled, and the high clouds are grayish or have a uniform layer with a pale bluish color. Turbulent clouds in the lower layer are dark gray thick clouds that extend from low to high, completely covering the sun and moon and making it rain.

TABLE. I. 10 CLOUD TYPES DEFINED BY THE WORLD METEOROLOGICAL ORGANIZATION

Name	Cloud bottom height (km)	Cloud top height (km)
Cirrus	6	
Cirrocumulus	6	
Cirrostratus	6	
Altocumulus	2	6
Altostratus	2	6
Nimbostratus	0	2
Stratocumulus	0	2
Stratus	0	2
Cumulus	0.5	6
Cumulonimbus	0.5	6

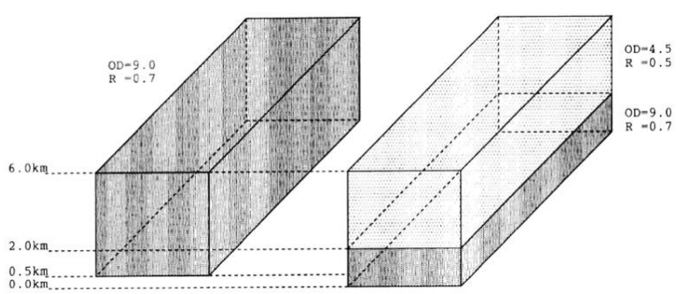


Fig. 1. Assumed One and Two Layer Clouds.

Cumulus clouds are clouds that appear about 2000 m from the ground, and are layered clouds such as gray or whitish or grayish white clusters or thin plates. The stratus has a uniform cloud bottom and drizzle with gray clouds. Such middle-layer clouds are optically thinner than the lower-layer clouds and pass solar radiation well. According to the GLAS model, the albedo of the high clouds (As) is 0.5 and the albedo of the stratus St is 0.7. Therefore, also from the GLAS model of [13], the upper cloud reflectance (R) is 0.5, the optical thickness (OD) is 4.5, the lower cloud reflectance (R) is 0.7, and the optical thickness. (OD) was set to 9.0.

III. MONTE CARLO SIMULATION IN THE PRESENCE OF CLOUDS IN A UNIFORM ATMOSPHERE

A. Simulation Outline

Numerically solving the problems of multiple scattering and reflection of light in the atmospheric and surface systems close to reality is considered by the Monte Carlo method. As shown in Fig. 2, assuming a three-dimensional atmospheric cell with an isolated cloud, photons are randomly generated from the upper surface of the atmosphere and incident so as to follow the solar incident direction (zenith angle and azimuth) and (μ_o, φ_o). Then, the travel distance of each generated photon and its scattering and absorption with atmospheric particles are tracked with probabilistic determination according to the atmospheric optical characteristic parameters.

In this simulation, the radiance emitted from the upper surface of the atmosphere is calculated by counting the number of photons emitted from the upper surface of the atmospheric cell.

Here, the following assumptions are made.

When a photon leaves the range of the atmospheric cell due to scattering, the photon enters the atmospheric cell from a surface opposite to the surface.

The atmosphere is optically uniform within the same layer (uniform in the horizontal direction and non-uniform in the vertical direction).

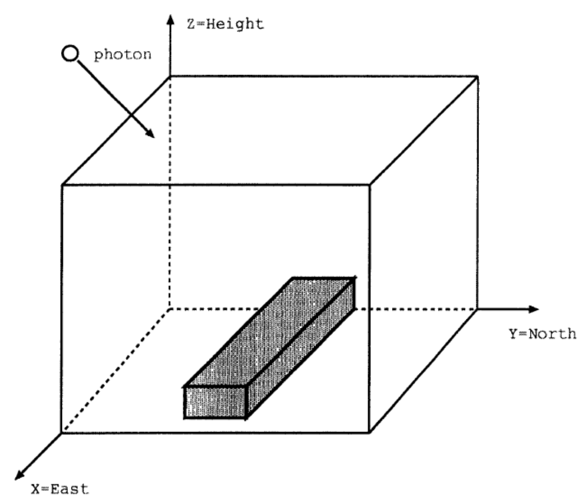


Fig. 2. Assumed a Three-Dimensional Atmospheric Cell with an Isolated Cloud.

The ground surface is a Lambertian surface that is flat and has a uniform reflectivity.

A photon enters the atmospheric cell according to the incident angle of the sun. At this time, the distance (free path) that the photon travels in the atmosphere is determined by using a random number, and it is determined according to their optical thickness whether the particle that the photon collides next is an air molecule or an aerosol particle. Determine whether the photon is absorbed or scattered by the single scattering albedo of the particle, and if it is scattered, determine the scattering angle by the phase function of the particle and also calculate the free path.

These are repeated, and photons propagate in the atmosphere. The surface of the cloud was assumed to be a Lambertian surface[3], reflected according to its reflectivity, and transmitted with the transmittance minus the reflectivity. Furthermore, assuming that the ground surface is also a Lambertian surface, reflection was performed according to the reflectivity, and photons were absorbed with the absorptivity subtracted from 1.

Real clouds and the ground surface are thought to be surfaces with anisotropic two-way reflection characteristics such as Minnaert's law[4], but in particular, there was no paper mentioning reflection characteristics such as the minerality coefficient of clouds. Here, the assumption of Lambert property was also made from the ease of calculation. For photons that jump out of the simulation cell, the photons are incident in the same direction from the position opposite to the position where they jump out. The details of these Monte Carlo methods are described in [5].

In the Monte Carlo method, the number of photons incident on the atmospheric cell needs to be sufficiently large. When this is small, the variation of the solution is large and the simulation is in a state where the signal-to-noise ratio is bad. As an example, calculating the radiance 1 at the top of the atmosphere while changing the number of photons gives Fig. 3. This shows that about 1 billion photons are required to obtain stable atmospheric top radiance. Therefore, the number of photons used in the subsequent Monte Carlo method was set to 1 billion.

B. Simulation Parameters

The setting of the Monte Carlo simulation parameters was the same as in [5]. That is, the size of the atmospheric cell was 50km from east to west, 50 km from north to south, and 50 km high. The atmospheric composition was only air molecules and aerosols, and those contained in the extra-atmospheric solar radiant flux (MODTRAN[5]-3.7[14] were used.

The real part of the complex refractive index[6] of the aerosol was 1.44, the imaginary part was -0.005, and the single scattering albedo[7] was 0.9318. In addition, the aerosol particle size distribution is the Junge distribution[8], the Junge parameter is 3, the wavelength is 0.5 μm, the solar zenith angle is 30 degrees, the solar azimuth angle is 120 degrees from north to west, and the ground surface is uniform on the Lambert surface. When sunlight is incident from the 180 degree direction, the adjacent effect is the largest, and the incidence from the 90 degree direction is the smallest.

The quantitative relationship between the incident condition of sunlight and the adjacent effect will be described in another opportunity. Here, the case of 120 degree incidence is taken up as an example. In addition, the satellite altitude can be set arbitrarily, but here we assume the 500 km adopted by the JERS-1[9] satellite, etc. and the instantaneous visual field is assumed to be 700m × 700m on the ground surface. That is, the situation where a push bloom type radiometer with an instantaneous field of view of about 0.0014 radians = 0.8 degrees is mounted on the satellite is simulated. A bin of approximately 0.8 degrees x 0.8 degrees was set at the top of the atmosphere at 50 km altitude, and the number of photons entering the bin within the instantaneous viewing angle was counted.

The bins were arranged in the *x* direction (east-west) and scanned in the y direction (north-south) to simulate a push bloom. As shown in Fig. 4, the cloud was a rectangular parallelepiped (10 km in the x direction, 50 km in the y direction) and placed in the center of the atmospheric cell. The cloud surface was a Lambertian surface. The ground surface other than this cloud assumes a flat sea surface.

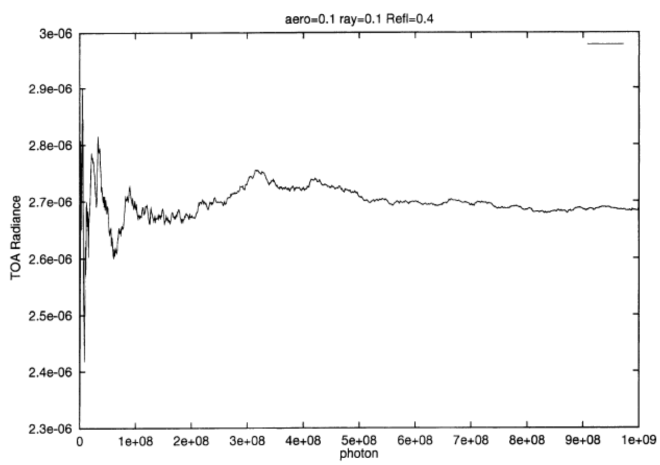


Fig. 3. A Relationship between the Top of the Atmosphere (ToA) Radiance and the Number of Photons[10] aero : Aerosol optical Depth, ray: Rayleigh Optical Depth, Refl : Reflectance.

C. Adjacency Effect

Differences in the adjacent effects of the one layer and the two layer clouds comes as evaluation items of the adjacent effects. Effects of the two-layer clouds on the adjacent effects

[3] https://en.wikipedia.org/wiki/Lambertian_reflectance

[4] https://www.researchgate.net/publication/228399602_Comparison_of_the_Minnaert_constant_for_different_forest_types_using_multi-temporal_SPOTHRV_data

[5] http://modtran.spectral.com/

[6] http://eodg.atm.ox.ac.uk/ARIA/

[7] https://en.wikipedia.org/wiki/Single-scattering_albedo

[8] https://sites.google.com/site/aerosolpedia/yong-yurisuto/guang-san-luan/10

[9] https://ssl.jspacesystems.or.jp/ersdac/Projects/JERS1/JOPS/JOPS_E.html

[10] The radiance when the ground surface reflectance is set to 0.4, the aerosol optical thickness is set to 0.1, and the air molecule optical thickness is set to 0.1.

of the clouds, effects on the side effects of the cloud, and adjacency of the phase function of the cloud particles. The effect on the effect was considered. Author in [5] clarified the difference in the adjacent effect between the first and second clouds, the effect of the two clouds on the adjacent effect by layer, the effect on the adjacent effect on the side of the cloud, etc. The effect of the phase function[11] of cloud particles is investigated.

In order to evaluate this adjacency effect, the slope factor (D99) defined in [5] is defined again. As shown in Fig. 5, D99 is the distance from the edge of the cloud until the influence of the adjacent effect on the brightness of the cloud gradually decreases until it reaches 1 % of the average sea surface brightness. This 1 % adjacency effect is equivalent to a few levels if the quantization bit number of the satellite-borne sensor is 8 bits, and cannot be ignored. Considering that the maximum input radiance of the onboard sensor is set so as not to saturate even when observing objects with high brightness levels such as clouds and snow and ice, this 1 % adjacent effect is reflected in the sea surface emission brightness. It is comparable and cannot be ignored.

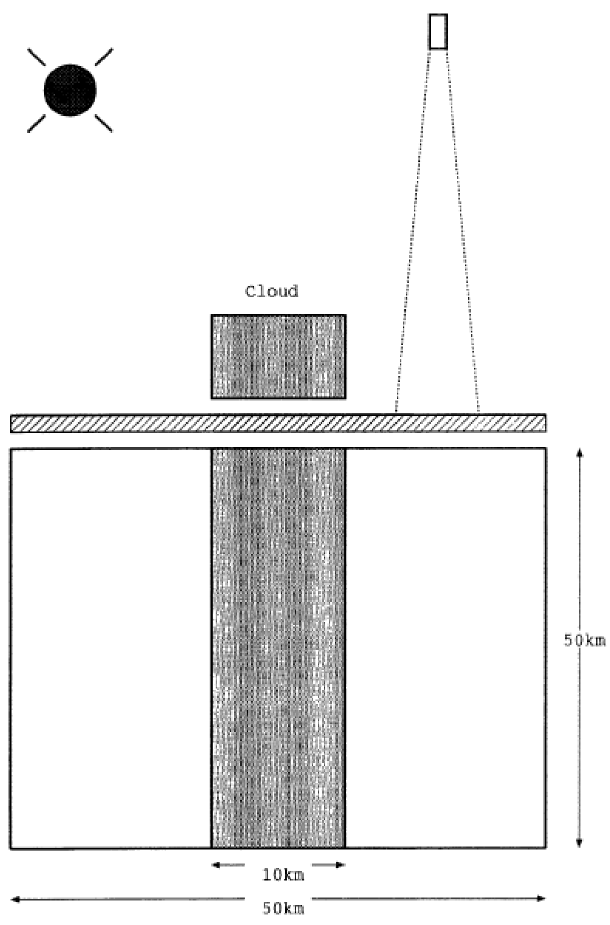


Fig. 4. Rectangular Parallelepiped (10 km in the x Direction, 50 km in the Y Direction) Cloud.

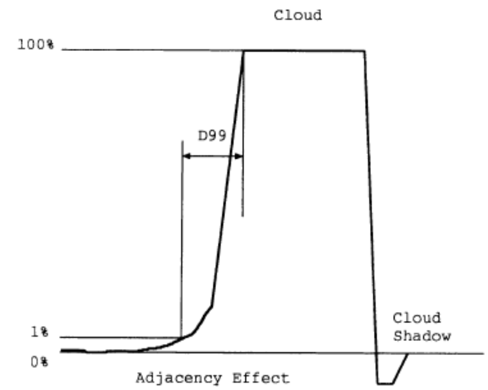


Fig. 5. The Definition of Slope Factor (D99).

IV. Experiments

A. Comparison between One Layer and Two Layer Clouds

Fig. 6 shows a comparison of phase functions between aerosol and cloud particles calculated with Mie scattering theory while Fig. 7 shows a comparison of the adjacency effect of the single layered cloud with phase functions of aerosol and cloud particle(the unit of Top of the atmosphere: ToA radiance is W/cm^2/str/micrometer).

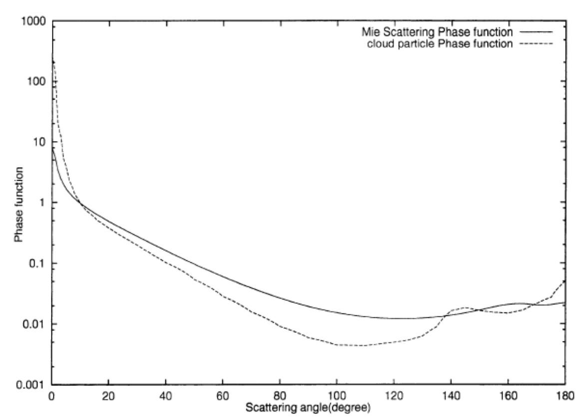


Fig. 6. A Comparison of Phase Functions between Aerosol and Cloud Particles Calculated with Mie Scattering Theory.

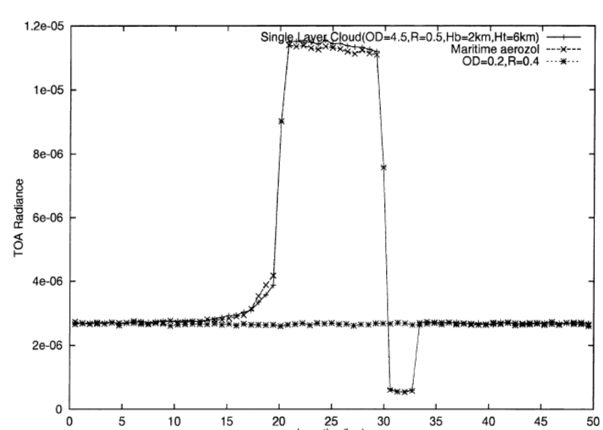


Fig. 7. A Comparison of the Adjacency Effect of the Single Layered Cloud with Phase Functions of Aerosol and Cloud Particle(the unit of ToA radiance is W/cm^2/str/micrometer).

[11] https://encyclopedia2.thefreedictionary.com/phase+function

The D99 of the first layer cloud is 4.8km and the D99 of the second layer cloud is 4.1km. The average radiance of the cloud is larger in the one-layer cloud than in the two-layer cloud. This is because the reflectance of the upper cloud of the two-layer cloud is 0.5, whereas the reflectance of the one-layer cloud is 0.7, and more reflection occurred on the upper surface of the one-layer cloud. It is done. In addition, because it is set so that sunlight enters from the left zenith angle 30 degrees, a shadow of about 3 km appears on the right side of the cloud on the opposite side, and the upper edge brightness at this point is rapidly decreasing.

Also, it is possible to show a comparison of the adjacent effects of clouds with the two-layer clouds and the upper clouds removed. The D99 of the two-layer cloud is 4.1 km, while the D99 of the one-layer cloud with the cloud above the two-layer cloud removed is 2.0 km, indicating that the adjacent effect is small.

Moreover, the cloud with the cloud height removed is lower than that of the two-layer cloud with a cloud top height of 2 km, and the cloud shadow is smaller. The cloud with the upper layer removed has a lower cloud top height and smaller volume than the one layer cloud, but the average atmospheric top brightness of the cloud is not much different. This shows that the upper atmospheric brightness of the cloud is dominated by the reflectance of the cloud upper surface.

The D99 of the two-layer cloud is 4.1 km, whereas the D99 of the one-layer cloud with the lower layer of the two-layer cloud removed is 7.6 km, indicating that the adjacent effect is considerably large. This shows that in the adjacent effect of two layers of clouds, the upper layer cloud increases the adjacent effect and the lower layer cloud decreases the adjacent effect. The cloud with the lower clouds removed has a greater adjacency effect despite the optical thickness of 4.5, reflectance of 0.5 and thin clouds. This is thought to be due to the fact that photons that are transmitted without being reflected by the cloud and scattered within the cloud have a great influence on the adjacent effect.

According to [5], the phase function of the cloud particle is an aerosol in the atmosphere (the complex refractive index is 1.44 for the real part, 0.005 for the imaginary part[12], and the aerosol distribution is the Junge distribution). The experiment was carried out assuming the same phase function as the Junge parameter 3, the minimum radius of the particle size was 0.01 μm, the maximum radius was 10 μm, and the wavelength was 0.5 μm. However, the actual cloud particle size has a maximum radius of 15 to 30 μm and the largest number is 3 to 7 μm [15], which is larger than the aerosol in the atmosphere.

For aerosols, the scattering theory of electromagnetic waves by particles of any size assuming a spherical shape, Mie scattering theory is used. In Mie scattering, when the wavelength of light is the same, the forward scattering becomes more prominent as the particle size of the scattered particles is larger. In this study, an average phase function based on the Mie scattering theory was obtained by considering the water soluble particle size of 15-30 μm in [15] as the phase function of cloud particles.

In the case of a single-layer cloud, the adjacent effect was compared when the phase function of cloud particles was the same as that of atmospheric aerosol and when the average cloud particle phase function was used. The results are shown in Fig. 7 and Table II. When the phase function of the cloud particle is that of an average cloud particle that increases forward scattering, the brightness near the cloud edge increases, but both the average radiance and D99 of the cloud decrease.

As shown in [5], an experiment was conducted on a cloud from which the lower cloud of a two-layer cloud that was relatively affected by the adjacent effect was removed. Fig. 7 shows that in the range of 3 to 4 km from the edge of the cloud, the adjacent effect is larger in the case of the cloud phase function than in the case of the aerosol phase function, and the ToA radiance is higher. This is thought to be due to the superior forward scattering of the phase function of the cloud particles than that of the aerosol. However, it turns out that there is almost no change with D99. Also, the shadow of the cloud appears at a distance as shown in Fig. 7 because the sun zenith angle is set at 30 degrees, and the brightness is the same in both cases.

The adjacent effect of the cloud is that the contribution of multiple scattered radiation in the cloud is large, so when the forward scattering of the phase function increases, the brightness near the cloud edge increases. It is found that the phase function of cloud droplets affects the adjacent effect [18].

B. *Monte Carlo Simulation in the Presence of Clouds in a Layered Atmosphere*

The real atmosphere is a non-uniform atmosphere in the altitude direction. Therefore, in order to assume a model close to reality, the three-dimensional atmospheric cell is made into a layer structure. This time, the atmosphere was divided by using Gauss-Seidel. Gauss-Seidel is a radiative transfer code based on a parallel plate atmospheric model using an iterative convergence calculation [5], [6]. The optical thickness is used to divide the atmosphere into several layers and calculate the radiative transfer. For an optical thickness of 0.2 (aerosol 0.1, air molecule 0.1), Gauss-Seidel divided the atmosphere into 10 layers. Table III shows the altitude and the optical thickness of aerosol and air molecules. For an optical thickness of 0.4 (aerosol 0.2, air molecule 0.2), the atmosphere was divided into 21 layers. Table IV shows the altitude and each optical thickness. Here, clouds were placed in these two optical-thickness layered atmospheric cells, and the effects of cloud adjacency were evaluated by the Monte Carlo method, which was more realistic.

TABLE. II. ToA RADIANCE AND D99 OF THE DIFFERENT PHASE FUNCTION OF AEROSOL AND CLOUD PARTICLE

Phase function	ToA Radiance	D99
Aerosol	1.099E-05	7.6km
Cloud particle	1.086E-05	7.6km

[12] This is equivalent to the mixture of aerosol derived from sea salt particles (Oceanic) and water-soluble aerosol (Water Soluble) in a ratio of 6: 4 according to the Maxell-Garnet mixing rule.

TABLE. III. OPTICAL DEPTH OF AEROSOL AND MOLECULE FOR TEN LAYERED ATMOSPHERE MODEL

Layer	Altitude(km)	Aerosol	Molecule
10th	14.868~50.0	0.007325	0.011809
9th	9.868~14.868	0.006168	0.013963
8th	6.668~9.868	0.00441	0.015632
7th	4.328~6.668	0.004286	0.0158
6tyh	2.708~4.328	0.006514	0.013617
5th	1.723~2.708	0.010541	0.009586
4th	1.089~1.723	0.013244	0.006767
3rd	0.663~1.089	0.014925	0.005178
2nd	0.282~0.663	0.015966	0.004171
1st	0.0~0.282	0.016615	0.003473

TABLE. IV. OPTICAL DEPTH OF AEROSOL AND MOLECULE FOR 21 LAYERED ATMOSPHERE MODEL

Layer	Altitude(km)	Aerosol	Molecule
21st	20.018~50.0	0.004726	0.010445
20th	15.568~20.018	0.008392	0.010709
19th	12.618~15.568	0.006839	0.012443
18th	10.418~12.618	0.005588	0.01373
17th	8.668~10.418	0.004682	0.014532
16th	7.188~8.668	0.004138	0.01501
15th	5.888~7.188	0.003708	0.015549
14th	4.788~5.888	0.004119	0.0151
13th	3.848~4.788	0.004819	0.014463
12th	3.068~3.848	0.006196	0.013183
11th	2.458~3.068	0.007987	0.011137
10th	1.978~2.458	0.010032	0.009337
9th	1.596~1.978	0.011502	0.00708
8th	1.284~1.596	0.012556	0.006634
7th	1.023~1.284	0.013349	0.005733
6tyh	0.798~1.023	0.014082	0.005081
5th	0.6~0.798	0.014738	0.004579
4th	0.426~0.6	0.015085	0.004109
3rd	0.27~0.426	0.015485	0.003754
2nd	0.129~0.27	0.015809	0.00345
1st	0.0~0.129	0.016158	0.003205

C. *Method for Calculation of the Parameters in the Structured Atmosphere*

The method of determining the photon travel distance at the boundary of layer structure with different optical properties cannot be used in the uniform atmosphere [16]. Therefore, when a photon enters an atmospheric layer with different optical properties, the photon is temporarily stopped at the boundary, and the traveling distance is obtained by a method of newly determining the traveling distance based on the optical parameters of the atmospheric layer newly entering at the boundary [17]. This method is shown in Fig. 8.

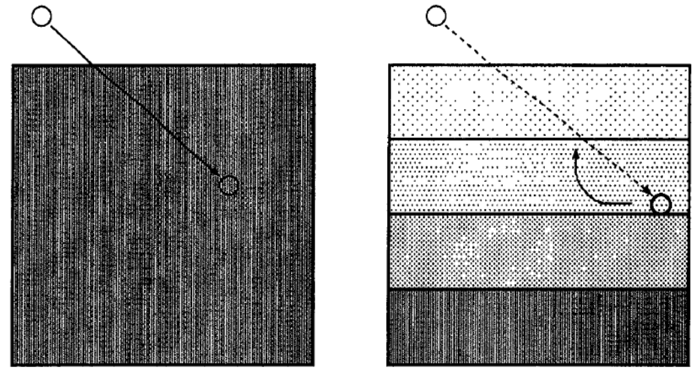


Fig. 8. The Method for Determination of the Final Destination of Photon when the Photons Across the Border of the different Layers.

D. *Limitation of the Proposed Simulation*

There is no limitation of the number of layers, cloud types as well as shape of the clouds for the proposed simulation method. It, however, is difficult to modelized nimbostratus due to the fact that physical characteristics of the nimbostratus is still unclear.

E. *Adjacent effects when a Single Cloud exists in the 10-Layer Atmosphere*

Fig. 9 and Table V show the adjacent effects when a single cloud exists in the 10-layer atmosphere.

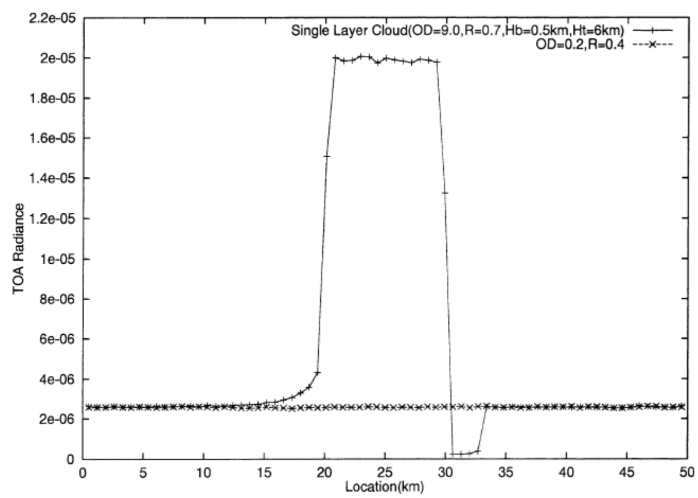


Fig. 9. Adjacent Effects when a Single Cloud Exists in the 10-Layer Atmosphere (A Comparison of the ToA Radiance for the Single and Ten Layered Atmosphere (the unit of ToA Radiance is $W/cm^2/str/Micrometer$)).

TABLE. V. A COMPARISON OF THE MEAN TOA RADIANCE AND D99 FOR THE SINGLE LAYERED CLOUD BETWEEN THE SINGLE AND THE 10 LAYERED ATMOSPHERE

Number of Layer	ToA Radiance	D99
10	1.91E-05	5.5km
1	1.54E-05	4.8km

Both the D99 and the average radiance of the cloud are larger in the one-layer cloud in the ten-layer atmosphere than in the one-layer cloud in the uniform atmosphere. Fig. 10, Table VI shows the adjacent effect when a single cloud exists in the 21 layer atmosphere. Similar to the 10-layer atmosphere, the one-layer cloud of the 21-layer atmosphere has a higher D99 and average cloud radiance. In the uniform atmosphere without clouds and the layered atmosphere, the top edge brightness of the atmosphere did not change much even when the reflectance of the ground surface was changed. There was a considerable difference between the effect and the average radiance of the clouds. This is because the upper surface of the cloud is in the 10th layer atmosphere, the 7th layer in the 10th layer atmosphere, and the 15th layer in the 21st layer atmosphere, and the 15th layer and the upper layer of the atmosphere. The actual atmosphere is considered to have a layered structure, and in order to investigate the influence of the adjacent effect more accurately, it is necessary to evaluate the adjacent effect in the layered atmosphere.

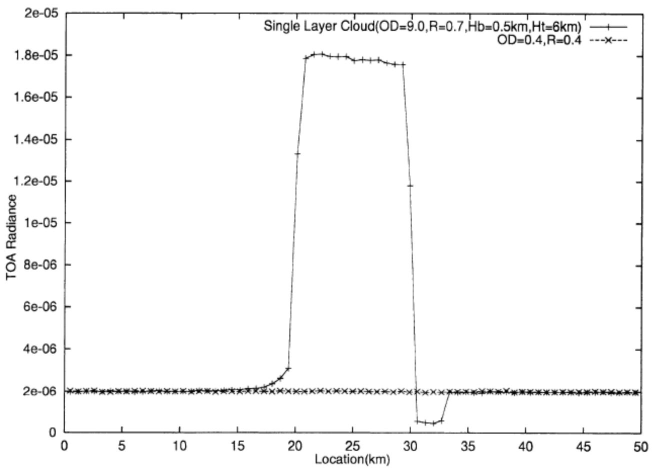


Fig. 10. Adjacency Effect of the Single Layered Cloud for the 21 Layered Atmosphere (the unit of ToA Radiance is W/cm^2/str/micrometer).

TABLE. VI. TOA RADIANCE AND D99 FOR THE SINGLE LAYERED CLOUD WITH THE SINGLE AND 21 LAYERED ATMOSPHERE

Number of Layer	ToA Radiance	D99
21	1.71E-05	3.4km
1	1.35E-05	2.7km

V. CONCLUSION

The adjacent effect of the cloud has a large contribution of radiation scattered in the cloud, so when the forward scattering of the phase function increases, the brightness near the cloud edge increases. This shows that the phase function of cloud particles needs to be considered when evaluating the cloud adjacency effect. In addition, from the evaluation results of the cloud adjacency effect when the number of layers is changed when the atmosphere is considered as a multilayer parallel plate model, it is confirmed that the adjacency effect increases as the number of layers increases.

In other words, since the adjacent effect is caused by multiple scattering in the atmosphere, the multilayer atmosphere is larger than the single layer, and an evaluation close to reality can be expected. This indicates that it is necessary to make the number of layers in the parallel plate atmosphere sufficiently large when evaluating the cloud adjacency effect.

VI. FUTURE RESEARCH WORKS

Further experimental studies are required for the validation of the proposed method. Also, applicability of the proposed method has to be confirmed through further experiments.

ACKNOWLEDGMENT

The author would like to thank Prof. Dr. Hiroshi Okumura for his valuable discussions.

REFERENCES

[1] Kawata, Y., Y. Haba, T. Kusaka, S. Ueno and Y. Terashita, Atmopsheric effects and their cor-rection in airborne sensor and Landsat MSS data,cProceedings of the 12th International Symposium on Remote Sensing of Environment, ERIM, pp.1241-1257, 1978.

[2] Lenoble, J., Edt., Radiometric transfer in scattering and absorbing atmospheres: Standard computational procedures, A Deepack, Publishing 1985.

[3] Tsutomu Takashima and Kazuhiko Masuda, Algorithm for calculating the influence of land reflected light on the radiance from the top of the atmosphere in coastal areas, Journal of the Remote Sensing Society of Japan, Vol.17, No.4, pp.3-13, 1997.

[4] Wilson A., Atmospheric correction of SeaWiFS data: An algorithm and validation for case-2 waters, NASA Technical Memorandom 104566, 31, 1996.

[5] Kobayashi, T., K. Masuda and M. Sasaki, Effects od adjacent clouds to satellite field of view on the measurements of the surface refclectivity, Proceedings of the International Symposium on the Atmospheric Correction of Satellite Data and Its Application to Global Environment, 157-160, 1998.

[6] Kohei Arai and Shinya Kawaguchi, Evaluation of peripheral effects considering cloud layer structure by Monte Carlo method, Journal of Remote Sensing Society of Japan, Vol.21, No.2, pp.179-185, 2001.

[7] Arai, K., Terayama, K., Evaluation of the effects of the peripheral effects of clouds with a layer structure based on the Monte Carlo method, Chiba University Research Center for Environmental Remote Sensing, No.6, pp.52-55, 2000.

[8] Kawada, T., Evaluation of the influence of surrounding clouds on target observation brightness by Monte Carlo method, No. 6, pp. 29-33, 2000.

[9] Kohei Arai, Adjacency effect of layered clouds estimated with Monte-Carlo simulation, Advances in Space Research, Vol.29, No.19, 1807-1812, 2002.

[10] Kohei Arai, Masanori Sakashita, Evaluation of Cirrus Cloud Detection Accuracy of GOSAT/CAI and Landsat-8 with laser Radar: Lidar and Confirmation with Calipso Data, International Journal of Advanced Research on Artificial Intelligence, 5, 1, 14-21, 2016.

[11] Kohei Arai, Masanori Sakashita, Hiroshi okumura, Shuichi Kawakami, Kei Shiomi, Hirofumi Ohyama, Comparative Study on Cloud Parameter Estimation AmongGOSAT/CAI, MODIS, CALIPSO/CALIOP and Landsat-8/OLI with Laser Radar as Truth Data, International Journal of Advanced Research on Artificial Intelligence, 5, 5, 21-29, 2016.

[12] Kohei Arai, Thresholding Based Method for Rain, Cloud Detection with NOAA/AVHRR Data by Means of Jacobi Itteration Method, International Journal of Advanced Research on Artificial Intelligence, 5, 6, 21-27, 2016.

[13] Kohei Arai, Shinya Kawaguchi, Elucidation of adjacent effects of multilayer clouds based on Monte Carlo method, The 30th Annual Conference of the Remote Sensing Society of Japan, B8, pp.49-50, 2001.

[14] WE. Meador, W.R. Weaver, Two-Stream Approximations to Radiative Transfer in Planetary Atmospheres: A Unified Description of Existing Methods and a New Improvement, Journal of Atmospheric Science, 37, 1980.

[15] Tatsushi Tokioka, Masaki Yamasaki, Nobuo Sato, Numerical simulation of meteorology (Meteorological Class 5). The University of Tokyo Press, 247 pp., 1993.

[16] US Air Force Geophysical Laboratory, MODTRAN 3 User Instructions, GL-TR-89-0122, 1996.

[17] Tomio Asai, Takeda Ikuo, Kimura Ryuji: "Atmosphere Science Course 2 Atmosphere with Clouds and Precipitation" (University of Tokyo Press, p.93., 1981.

[18] Arai Kohei, Basic Theory of Remote Sensing, Academic Book Publishing, 2001.

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is 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 37 books and published 570 journal papers. He received 30 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.

An Intelligent and Adaptive Model for Change Management

Ali M. Alshahrani
Arab Open University (AOU)
Riyadh 11681, Saudi Arabia

***Abstract*—The continuous and rapid changes that are taking place in the world today makes the change management crucial to any organization. The existing change management models bridge the gap between motivation, planning, and implementation. These models are highly significant as the organizational change is not a rare event anymore, but it is an ongoing process, and the 'business as usual' model becomes insignificant to most organizations. To automatize the theoretical model of change management, an intelligent and adaptive model for change management is developed in this paper, which takes into consideration all the positive and negative effects (factors) that may take place at any time and any place internally (internal factors) or externally (external factors). Based on these factors, accordingly, the proposed model can efficiently find a reasonable solution that adapts to the existing situation to avoid any failure of organizational management. The proposed system is built based on a decision support system (DSS) with inputs that represent the influencing factors and an output that represents feedback on the method of management. In this paper, the proposed change management model has been verified, and the results have been reported accordingly.**

Keywords—Change Management; External Environment; Internal Environment; Decision Support System (DSS)

I. Introduction

Change management is a transformation model that deals with the changing elements in an organization. Vlados [1] defined change management as "the forms and ways to design, implement, control, evaluate, and assimilate changes". In the business management field, significant changes should be implemented as a response to the changing elements in an organization [2], which are carried out using the change management models. With the rapid changes in the world nowadays, change management becomes crucial to every organization. Generally, a change management model comprises a set of processes that are required to achieve the desired outcome and to deal with the management aspects that are affected by the change. In terms of elements, change management concerns about people, processes, and culture [3]. Change management also covers other related elements, such as the strategic direction and the personal development of the staff [4]. In terms of process, change management implements planning, initiating, realizing, controlling, and finally stabilizing the change processes on both, corporate and personal level.

Although change management has been around for a while, it becomes more popular recently for its role in initiating significant changes to work procedure and culture in rapidly changing fields, such as pollution prevention, fertilizer, water resource, energy, raw materials management, pesticide residues, etc. [5]. According to Graetz [6] "increasing globalization, deregulation, the rapid pace of technological innovation, a growing knowledge workforce, and shifting social and demographic trends, few would dispute that the primary task for management today is the leadership of organizational change". Accordingly, the needs for change management approach is increasing nowadays like no time before. The organizational change is not a rare event nowadays, but it is an ongoing process, which makes the models of 'business as usual' becomes insignificant to most organizations. The demands for change management models are increasing day by day, and the developed approaches are filling this gap with inventions of various models.

Various models for change management have been developed to address the change in the environment controlling the organizations. The current researches have created formal environmental management models. Earlier models for change management, such as the awareness, desire, knowledge, ability and reinforcement (ADKAR) model, Kotter's model, Bridge's model, McKinsey's model and Lewin's model [7], focused on specific feature, element or process, such as the planning, human adaptation to change or the theoretical characteristics of the change management process. While the recent approaches that are developed for change management focus on analyzing the factors that influence the change processes, moreover, these approaches provide suitable frameworks to be adopted in the implementation of the change based on these factors [8]. Accordingly, a set of predetermined factors are established, these factors are monitored, and the suitable management method is then adapted based on these factors. Overall, these models bridge the gap between motivation, planning, and implementation. These models are highly significant as the organizational change is not a rare event anymore, but it is an ongoing process, and the 'business as usual' model becomes insignificant to most organizations.

Kolk and Mauser [9] gave an overview of the process by which the development of such environmental management models can be implemented. Besides, model analysis and model evaluation in term of characteristics, strengths, and weaknesses have been discussed and clarified. Although the change management models have been studied extensively, automatic approaches for change management is still immature. The computational provision of the change management models is limited and of rule-based basis. Rule-based depends on inputs based on the pre-determined change factors and selects among pre-determined actions encoded in the rules as the proper response for the change factors. Although the rule-based approach is easy to be interpreted by the user, it is erroneous as it required accurate rules based on the human expertise, which could not be accommodated when the changes are taken places rapidly. Accordingly, intelligent change management model needs to be developed based on machine learning approach.

In this paper, an intelligent and adaptive model for change management is developed based on a decision support system (DSS) with inputs that represent the influencing factors and an output that represents feedback on the method of management. The proposed approach takes into consideration all the positive and negative effects (factors) that may take place at any time and any place internally (internal factors) or externally (external factors). In this way, change management is carried out using a practical way immediately as responding to any changes in external and internal environmental changes. Using this approach has been hypothesized to ensure that the correct selection is guaranteed. The rest of this paper is organized as follows: Section 2 presents the related work. Section 3 presents the factors of the change management. In Section 4, the details of the proposed approach are described. Section 5 is devoted to the evaluation, and finally, Section 6 is a conclusion.

II. Related Work

Atomization the business process management is critical to provide semantics to the business process and systemize the process. In this aspect, Figueiredo and de Oliveira [10] used ontology to represent the business process. According to Figueiredo and de Oliveira [10] "ontologies conceptualize and organize the information that is embedded and unstructured in the business processes and that must be explored. They structure the implicit knowledge that is present in the business processes, enabling the understanding of this knowledge by machine". Accordingly, the business process can be analyzed automatically based on the utilized ontology, anomalies can be detected and the changes can be conducted manually based on the detected changes.

Sarno and Sinaga [11] used ontology to capture the business process anomalies in comparison with the company's principles. Unlike the approach that is proposed by Figueiredo and de Oliveira [10], Sarno and Sinaga [11] approach detects anomalies automatically. Yet, the change management is implemented manually based on the captured anomalies. Ariouat et al. [12] developed an ontology for business process management based on three perspectives, these are behavioural organizational, and social. Rules are defined to deduce knowledge from the constructed ontology and changes can be deduced in similar ways. Similar other approaches based on ontology and knoweldge extraction have been proposed to autmatize the buisness process [13, 14]. Yet, these systems required human intervention to response to the changes.

Oriol et al. [15] proposed a rule-based model for business process management with constraints violation checking. The input to the model are the business process that are represented in UML, and the output is repair activity. Overall, based on the reviewed literature, the change management process is implemented as a manual step after analyzing the business process.

The computational provision of the change management models is limited. The existing change management models can be characterized as a rule-based approach that derived suitable actions based on a set of criteria. The developed approaches used different factors and implemented different actions. Ayhan et al. [16] proposed a multi-agent system for change management, in which agents are monitoring the changes in an organization, sharing the change information and cooperating their response. Agents are responsible for identifying the change factors, such as the capabilities of the organization to change and selects among pre-determined actions, the proper response based on the change factors.

III. Change Management Factors

In order to quantify the factors to be considered in the change management models, Kuipers et al. [17] reviewed the existing literature on the change management and identified four factors by which change is influenced and through which the change management is taking place, these are context, content, process, and outcomes.

The context refers to the internal and external factors, such as policies, political and economic changes. As a response to the context, Macleod and By [18] studied the effects of the context on the change management of public sector organizations and analyzed the benefits of using historical context data in make decisions about the management techniques in public sectors. Similarly, Modell [19] addressed the context influence in a healthcare organization. While De Boer et al. [20] studied the context of educational change management. These studies illustrated the influence and benefits of the context using use-cases with context-related data analysis. The content refers to the strategies of the organization, such as the management structure and hierarchy [21]. The existing literature on content-based change management focused on the effects of policies in public sectors that are related to the public organization. Schout [22] discussed and criticized the existing change management techniques in relation with the change in policies and regulations within public sector organization and highlighted the benefits of reflecting the change management based on the details change in the policies as parallel to using the historical data in change management technique. The processes refer to the periodically organizational process, duties, and roles. Erakovic and Powell [23] proposed three ways for change management based on the process influence; these are incremental change management, radial and reductive. The outcomes refer to the experiences gained by the change and which should contribute to the management of future changes [24]. A literature review on the change management outcomes showed that the outcomes of the change management are four types, these are readiness for change, commitment to change, openness to change and cynicism about organizational change [25]. Various studies have focused on change management concerning the outcome. Paul Battaglio Jr and Condrey [21] studied the change management from an outcome perspective in the public sector, while Lindquist [26] studied the outcome in change management in the private sector.

Overall, change management in public, and private sectors are equal in recent studies [27, 28]. The theoretical models for change management are presented in the literature, yet there is a need to develop an automatic change management technique for general purpose based on analyzing and processing previous data from the factors above.

IV. The Proposed Approach

In order to develop a comprehensive environmental management model in a rapidly changing environment, a data mining approach based on DSS is proposed in this paper. The proposed intelligent and adaptive model for management in the rapidly changing environment uses various components that are integrated to produce consistent and acceptable decisions as given in Fig. 1, these components are the external environment factors, the internal environment factors, decision support system and the intelligent change management process. The proposed model is developed by determining the influenced factors, then, identifying the DSS resources and finally setting the overall process by which the decision-making is implemented.

A. External and Internal Environmental Factors

Any organization is related to a community and the surrounding environment from which it obtains the resources and to which it supplies products and services. In details, without the financial recourses and human resources, the organization seize to exist. Moreover, an organization has to obey and follows legalizations and other requirements imposed by the local government, as well as following the trends of the market and competitors. Besides, organizations must respond to any political change inside or outside the country. Accordingly, an organization is established based on a mutual relationship with its environment social, financial, regulations and market status and conditions. Overall, the factors of the external environment, as listed in Table I, which are considered by the proposed model are the political, social and organizational bylaws, the economic, competition, and investment requirements and the environment and ecology requirements. The internal environment factors are human resources, budget and financial situation, and the products/ services factor.

B. Decision Support System (DSS) Resources

The DSS of the proposed approach is built based on the following resources: the organizational database that contains information about the internal factors, information resources for the external factors, which are obtained using information retrieval from the internet, planning studies and marketing studies, as listed in Table II. The information systems play a significant role to perceive the environmental changes and to act in the right way. Information systems are used to collect environmental data and to identify the external and internal changes that urge the organization response. New coming technologies, new products and changing of the public tastes and values put strains on an organization's culture, politics, and people. The inertia built into an organization's standard operating procedures, the political conflict raised by changes to the existing order, and the threat to closely held cultural values typically inhabit organizations from making significant changes. Environments usually change much faster than organizations. Accordingly, the organization has to adapt to a rapidly changing environment by continuously collecting information about the surrounding environment.

Accordingly, in the proposed system, the external factors, which are determined in the previous subsection, are controlled by a set of search terms, and statistics about the retrieval information from the internet are saved and used for decision-making. The values of the internal factors are managed internally by the organization and the saved in the database system, which is accessed by the DSS. The marketing and planning studies are represented as another set of factors. All together forms a timely changed feature vector that is used to make the suitable action and adapt to the environmental change in the proposed change management system, as illustrated in the example given in Table III.

TABLE. I. External and Internal Environmental Factors

Index	**Factor**	**Index**	**Factor**
FM1	Political	FM6	Investment requirements
FM2	Social	FM7	Environment and Ecology
FM3	Organization bylaws	FM8	Human resources
FM4	Economic	FM9	Budget & financial situation
FM5	Competition	FM10	Products/services

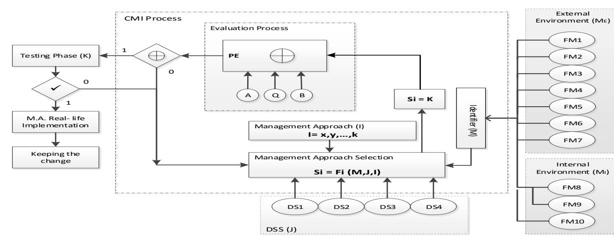


Fig. 1. An Intelligent and Adaptive Model for Change Management.

TABLE. II. Decision Support Decision Elements

Index	**Element**
DS1	Internet (For external factors)
DS2	Database (For internal factors)
DS3	Marketing Studies
DS4	Planning Studies

TABLE. III. Example Feature Vector

Factor	**Change**
Political	Minor Change
Social	No Change
Bylaws	No Change
Economic	Major Change
Competition	No Change
Investment	Minor Change
Environment	No Change
Human	No Change
Budget	Major Change
Product	No Change
Marketing	No Change
Planning	No Change

C. Model Description

The cooperation of the previous factors leads to determine the decision to be made for change management, which put the organization on the right path and avoids any failure of the business. The proposed model for change management process inputs, which are represented by the discussed factors, and produce the outputs as a decision to be made. The input variables are aggregated in a hierarchy of factors with three levels; the bottom level involves all the features discussed previously. In the second level, the variable *M* combines all the external and internal factors, which are: political factor, social factor, organizational bylaws factor, economic factor, competition factor, investment factor, and environment factor. Besides, this variable combines the internal factors, which are human resources, budget and product/service factors. Variable *J* combines the following: planning studies, marketing studies and databases. Finally, the variable *I* refer to the method(s) of management, which can be centralized, distributed, etc.

The processing stage of the system includes a management method (approach) selection and evaluation process. The management approach selection step used the inputs, which are represented as *M*, *J* and *I* and to select the best method that will be suitable to a current situation of organization taking into account based on the inside and the outside effects. An intermediate result is that the management selection step will go through the evaluation process. The evaluation process is a subsystem that measures and evaluates the selected management approach based on AQB criteria, which are:

- A: stands for the employees' level of acceptance to apply a new management approach measured in percentage.
- Q: stands for quality measurement, and it refers to the percentage by which the new management approach can meet the quality standard requirements in a given organization.
- B: stands for a budget of an organization and refers to the percentage by which the budget can support the selected approach of management.

The output of the evaluation process can be negative or positive. In a negative case, a loopback step will be carried out to select another management approach while in a positive case, the selected approach for change management will go through a testing stage (phase) for a while to ensure that the selected approach is a right choice otherwise; the loopback process will take place. After a positive result of the testing phase, a select approach of organizational management will be confirmed.

V. Evaluation and Results

In the evaluation process, an online survey is conducted to establish a dataset that is used to evaluate to what extent the proposed model can fit with the change management in different organization. The target organizations were experts who are working in the education field, a telecommunication corporation and banking. Although these fields do not cover all organizations that require change management, we believe the results can be used to validate the proposed approach. The survey involves the representation of the variables M, J, and I with their related factors. Content Validity Index (CVI) is used for model evaluation [29]. The number of experts that are considered, according to the literature is 2-20, in this study 18 experts were contributed to the validation of the model [29]. Before filling in the survey, a brief on the developed model was given to the experts. CVI is calculated as the percentage of scores of 3 or 4 to the total number of scores (number of questions in the survey). Lynn [30] suggests that above 0.78 is accurate results with this number of experts. The complete questioner is given in Appendix A. The results of the survey are given in Table IV.

Out of the 18 raters, 8 raters rates the correctness and the applicability with scale 4. The rest of the raters have awarded mixed of scale 3 and scale for the two criteria. Helpfulness comes in the second order among criteria. All the rates have rated the helpfulness with mixed of scale 2, scale 3 and scale 4. The low rates were awarded for attractiveness and easiness. To justify this issue, it was realized that the rates does not know what DSS is and how such system could be implemented. Accordingly, easiness can only be evaluated with a full empirical study in the future.

As for the testing phase, the experiments depend on the following factors:

- Human resources factor (A) that measures the acceptance rate of the new management approach among human resources in an organization.
- Budget and economic situation of an organization factor (B), which involves questioning about whether the budget and economic situation can support and endure the consequences of applying a new management approach.
- Quality factor (Q), a new approach of management should achieve minimum requirements of quality issues within an organization.

To explain the evaluation process that is proposed, the system supposes represent each of the factors, A, Q and B using 2-bit code and the combination of these bits represents the percentage of the effects on the change management process. This percentage reflects the performance of the system for change management as the following:

- 00: 0%
- 01: 25%
- 10: 50%
- 11: ≥ 75%

The results of the evaluation process based on a constructed dataset based on the survey results are conducted and after careful analysis, it was concluded that for the input factors with a minimum value of 50% for B and Q while a minimum of 0% for the A factor, in some particular cases lead to acceptance of the management style being followed. Table V illustrates the AQB factors presented in a combination of 6 bits (2 bits for each factor) along with results of the evaluation process in a proposed model.

TABLE. IV. SURVEY RESULTS

Element	Value
Number of raters	18
Average pairwise percentage	70%
Total number of answers	18 (Experts) *15 (Questions) = 270
Total number of scale 1 answers	28
Total number of scale 2 answers	30
Total number of scale 3 answers	122
Total number of scale 4 answers	90
CVI	**0.785**

TABLE. V. EVALUATION RESULTS

AQB	Output	Results
000000	0	x
000001	0	x
……..	……..	……..
001011	1	✓
001010	1	✓
……..	……..	……..
100000	0	x
……..	……..	……..
101010	1	✓
111111	1	✓

VI. CONCLUSION

In this paper, a new intelligent and adaptive model for change management has been developed to carry out change management based on a set of influencing factors and a decision-making process that uses a decision support system (DSS). All the internal and external factors of the change management were taken into account. The output of the proposed model is either a negative or a positive indication. In a negative case, a loopback step can be carried out to select another management approach while in a positive case, the selected approach for change management will go through a testing stage (phase) for a while to ensure that the selected approach is a right choice otherwise. The pilot experiments of the proposed model show its effectiveness in identifying the correct/wrong decision for change management. The correctness and the applicability of the proposed approach were proved to be satisfied. Out of the 18 raters, 8 raters rates the correctness and the applicability with scale 4. The rest of the raters have awarded mixed of scale 3 and scale for the two criteria. The CVI of the pilot study was 0.78.

In the future, some cases will be experimentally implemented to show the efficiency of the model, an enhanced algorithm that explains the work of the proposed will also be developed. The goal of the future work will be enhancing the rates of the attractiveness and easiness criteria.

REFERENCES

[1] C. Vlados, Change Management and Innovation in the 'Living Organization': The Stra. Tech. Man Approach. Management Dynamics in the Knowledge Economy, vol. 7, no. 2, 2019, pp. 229-256.

[2] H. Kerzner and H.R. Kerzner, Project management: a systems approach to planning, scheduling, and controlling, John Wiley & Sons, 2017.

[3] M. Rosemann and J. vom Brocke, The six core elements of business process management, in Handbook on business process management, Springer, 2015, pp. 105-122.

[4] A. Alkhafaji and R.A. Nelson, Strategic management: formulation, implementation, and control in a dynamic environment, Routledge, 2013.

[5] Z. Malek, P. H. Verburg, I. R. Geijzendorffer, A. Bondeau, and W. Cramer, Global change effects on land management in the Mediterranean region, Global Environmental Change, vol. 50, 2018, pp. 238-254.

[6] F. Graetz, Strategic change leadership, Management decision, vol. 38, no. 8, 2000, pp. 550-564.

[7] S. T. Hussain, S. Lei, T. Akram, M. J. Haider, S. H. Hussain, and M. Ali, Kurt Lewin's change model: A critical review of the role of leadership and employee involvement in organizational change, Journal of Innovation & Knowledge, vol. 3, no., 3, 2018, pp. 123-127.

[8] S. Oreg, J. M. Bartunek, G. Lee, and B. Do, An affect-based model of recipients' responses to organizational change events, Academy of Management Review, vol. 43, no. 1, 2018, pp. 65-86.

[9] A. Kolk and A. Mauser, The evolution of environmental management: from stage models to performance evaluation, Business strategy and the environment, vol. 11, no. 1, 2002, pp. 14-31.

[10] L.R. Figueiredo and H.C. de Oliveira, Automatic Mapping of Business Process Models for Ontologies with an Associated Query System, International Conference on Enterprise Information Systems, 2018.

[11] R. Sarno and F.P. Sinaga, Business process anomaly detection using ontology-based process modelling and multi-level class association rule learning, International Conference on Computer, Control, Informatics and its Applications (IC3INA), 2015.

[12] H. Ariouat, C. Hanachi, E. Andonoff, and F. Benaben, A conceptual framework for social business process management, Procedia Computer Science, vol. 112, 2017, pp. 703-712.

[13] J. B. Gassen, J. Mendling, A. Bouzeghoub, L. H. Thom, and J. P. M. de Oliveira, An experiment on an ontology-based support approach for process modeling, Information and Software Technology, vol. 83, 2017. pp. 94-115.

[14] K. Kluza, K. Kaczor, G. J. Nalepa, and M. Ślażyński, Opportunities for business process semantization in open-source process execution environments, Federated Conference on Computer Science and Information Systems (FedCSIS), 2015.

[15] X. Oriol, G. De Giacomo, M. Estañol, and E. Teniente, Automatic business process model extension to repair constraint violations, International Conference on Service-Oriented Computing, 2019.

[16] M.B. Ayhan, M.E. Aydin, and E. Öztemel, A multi-agent based approach for change management in manufacturing enterprises, Journal of Intelligent Manufacturing, vol. 26, no. 5, 2015, pp. 975-988.

[17] B. S. Kuipers, M. Higgs, W. Kickert, L. Tummers, J. Grandia, and J. Van der Voet, The management of change in public organizations: A literature review, Public administration, vol. 92, no.1, 2014, pp. 1-20.

[18] C. Macleod and R.T. By, Organizational change management in public services: Key findings and emerging themes, Managing Organizational Change in Public Services, Routledge. p. 263-271, 2012.

[19] S. Modell, Performance measurement and institutional processes: a study of managerial responses to public sector reform, Management Accounting Research, vol. 12, no. 4, 2001, pp. 437-464.

[20] H.F. De Boer, J. Enders, and L. Leisyte, Public sector reform in Dutch higher education: The organizational transformation of the university, Public Administration, vol. 85, no. 1, 2007, pp. 27-46.

[21] R. Paul Battaglio Jr and S.E. Condrey, Reforming public management: Analyzing the impact of public service reform on organizational and managerial trust, Journal of Public Administration Research and Theory, vol. 19, no. 4, 2009, pp. 689-707.

[22] A. Schout, Organizational learning in the EU's multi-level governance system, Journal of European Public Policy, vol. 16, no. 8, 2009, pp. 1124-1144.

[23] L. Erakovic and M. Powell, Pathways of change: organizations in transition, Public Administration, vol. 85, no., 1, 2006, pp. 31-58.

[24] A.A. Armenakis and A.G. Bedeian, Organizational change: A review of theory and research in the 1990s. Journal of management, vol. 25, no., 3, 1999, pp. 293-315.

[25] Choi, M., Employees' attitudes toward organizational change: A literature review. Human Resource Management, vol. 50, no. 4, 2011, pp. 479-500.

[26] E., Lindquist, Organizing for policy implementation: The emergence and role of implementation units in policy design and oversight, Journal of Comparative Policy Analysis, vol. 8, no. 4, 2006, pp. 311-324.

[27] W. J. Kickert, Managing emergent and complex change: The case of dutch agencification, International Review of Administrative Sciences, vol. 76, no. 3, 2010, pp. 489-515.

[28] P. Klarner, G. Probst, and R. Soparnot, Organizational change capacity in public services: The case of the World Health Organization, Journal of Change Management, vol. 8, no. 1, 2008, pp. 57-72.

[29] R. K. Gable and M. B. Wolf, Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings, vol. 36, Springer Science & Business Media, 2012.

[30] M. R. Lynn, Determination and quantification of content validity, Nursing research, vol. 35, mo. 6, 1986, pp. 382-386.

APPENDIX A

The goal of this survey is to evaluate a model of change management through a set of variables and its related factors. The target of the survey is experts in education, telecommunication business, and banking. We hope that by conducting this survey, we will be able to capture your judgment/opinion about the model developed.

Instruction: Please state the extent to which you agree or disagree with the following statements (tick one per statement), where:

1 is Strongly disagree	**2 is Disagree**			
3 is Agree	**4 is Strongly agree**			
Q1. CORRECTNESS	SDIS	DIS	AG	SAG
	1	2	3	4
A. The variables of the proposed model are correctly identified	○	○	○	○
B. The factors related to each variable are correctly identified	○	○	○	○
C. The linkage between the factors and the variable are valid	○	○	○	○
D. The model is conceptually valid	○	○	○	○
Q2. APPLICABILITY	SDIS	DIS	AG	SAG
	1	2	3	4
E. The utilized variables can be used to guide the change management process	○	○	○	○
F. The factors that are considered cover all the features that influence the change management	○	○	○	○
G. The linkage between the factors and the variables can be used to model the change management factors	○	○	○	○
H. The model is applicable for any change management process	○	○	○	○
Q3. HELPFULNESS	SDIS	DIS	AG	SAG
	1	2	3	4
I. The model helps managers to evaluate the management decision in the change management process	○	○	○	○
J. The model helps managers to understand the factors that are considered for the change management process	○	○	○	○
Q4. ATTRACTIVENESS	SDIS	DIS	AG	SAG
	1	2	3	4
K. The model is widely applicable to evaluate the management decision in the change management process	○	○	○	○
L. The model is widely applicable to understand the factors that are considered for the change management process	○	○	○	○
M. The model is widely applicable to provide feedback about the organization change management process	○	○	○	○
Q5. EASYNESS	SDIS	DIS	AG	SAG
	1	2	3	4
N. The model is easy to be understood	○	○	○	○
O. The model is easy to be applied	○	○	○	○

A Systematic Review on Students' Engagement in Classroom: Indicators, Challenges and Computational Techniques

Latha Subramainan[1], Moamin A. Mahmoud[2]
College of Computing and Informatics
Universiti Tenaga Nasional
Kajang, Malaysia

***Abstract*—Students' engagement in a classroom is a key factor that influences several educational outcomes. Studies by the University of California, Los Angeles (UCLA) and British universities found that 40% of students are frequently experiencing boredom and less than 20% of students ask questions in class due to poor engagement. A survey by Malaysia's Program for International Student Assessment (PISA) found that 80% of the participating schools fell into the poor performance bracket. However, studies in this line of research are limited and scattered. To provide a clear insight into this problem and support researchers, it is crucial to understand the current state of research in this area. Consequently, in this paper, a comprehensive review is conducted to map the literature studies to a consistent taxonomy. Search terms revealed 87 papers from several databases that have been classified into seven categories. A systematic review method is applied, analysis is performed, and finally, findings, discussion, and recommendations are presented.**

Keywords—Classroom interaction; student engagement; engagement indicators; engagement challenges; computational techniques

I. Introduction

Over two decades, studies have reported that students' engagement is a crucial factor in predicting students' academic achievement and influencing several educational outcomes [1] [2] [3]. Several factors have been utilized by researcher to indicate engagement such as "self-report, attendance rates, teacher ratings, interviews, observations, cross-cultural data and assessments grades" [4] [5] [6]. On the other hand, researchers measure engagement by observing a student's active participation such as the amount of effort and positive emotion or via a student's voice and initiative to take personal responsibility for his/her behavior [6].

In this paper, we review the literature on students' engagement in a classroom. Search terms identified 87 papers from several databases such as Science Direct, Taylor and Francis, Sage, Springer, Wiley Online, IEEE Xplore. The collection of the selected papers is classified into seven categories, overview; dimensions and indications; research studies; problems; factors influence engagement; methods to measure engagement; and techniques to improve engagement. This paper starts by describing the method used to conduct this review. The method includes a search strategy, search terms, and selection process. We then review the seven categories. We also conduct a comprehensive analysis on the reviewed articles, which include data analysis on students' engagement research, number of selected articles by year of publication, rate of the reviewed articles in different categories, number of articles in different categories by year of publication, and taxonomy of the literature on student engagement. Three potential research problems are discussed, examining the impact of emotion on students' engagement; development of a configuration framework to generate an engagement strategy for a particular classroom environment; utilizing computer-based simulation and virtual environment to study and investigate poor engagement instead of using traditional research tools.

II. Method

A. Search Strategy

Six digital libraries and databases are selected to search for papers which are: "Web of Science, SagePub, IEEE Xplore, Springer, ACM Digital Library and Science Direct". The selection in this review includes those identified as relevant to school education, information technology, and social science.

B. Search Terms

The review aims to find all primary research work in conjunction with the terms for possible outcomes. The relevant search terms in the context of this review are identified as follows: "student-engagement in a classroom"; "student engagement in a classroom"; "student-involvement in a classroom"; "student involvement in a classroom"; "student-participation in a classroom"; "student participation in a classroom". Added to these terms are the keywords, "overview"; "research"; "studies"; "methods"; "status", "indications"; "problems".

C. Paper Selection Process

The following shows the stages of the selection process:

1) Search based on the search string.
2) Query (as discussed in Section 3.2).
3) Excluding duplicates/non-English articles.
4) Perform an abstract scan and exclude publications that do not address engagement research.
5) Full-text reading to filter and produce final set.

Using these conditions, 87 papers met the inclusion and exclusion criteria and are identified as relevant to the current review. These papers are classified into the following categories: an overview of students' engagement; dimensions and indications of students' engagement; studies conducted on students' engagement research; problems in students' engagement; factors influence students' engagement; methods to measure students' engagement; and techniques to improve students' engagement.

III. Analysis and Findings

During the analysis, we found 87 papers on students' engagement research. As shown in Fig. 2, nine papers on definition and history of students' engagement, 12 papers on dimensions of students' engagement, 13 papers on studies conducted on students' engagement research, 17 papers on problems in students' engagement, 10 papers on factors influence students' engagement, five papers on methods to measure students' engagement, and 21 papers on techniques to improve students' engagement.

Fig. 1 shows the number of reviewed articles in the seven categories according to the years of publication. The distribution from 1984 to 2018 is shown. The figure shows that, according to the reviewed articles, a few studies have been conducted between 1984 and 2003 at a rate of 12% of total reviewed papers. In the next five years, the rate increased to 21%. In the last decade, the rate reached 67%. This clearly shows that research in students' engagement is receiving increasing attention.

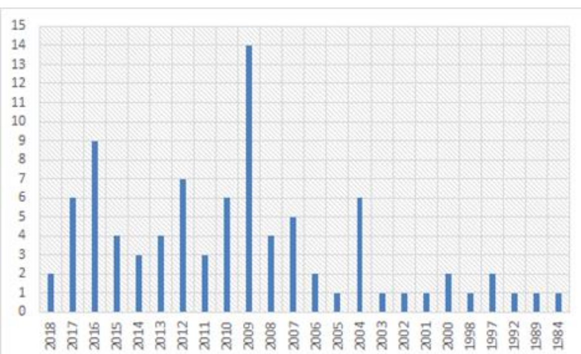


Fig. 1. Number of Selected Articles by Year of Publication.

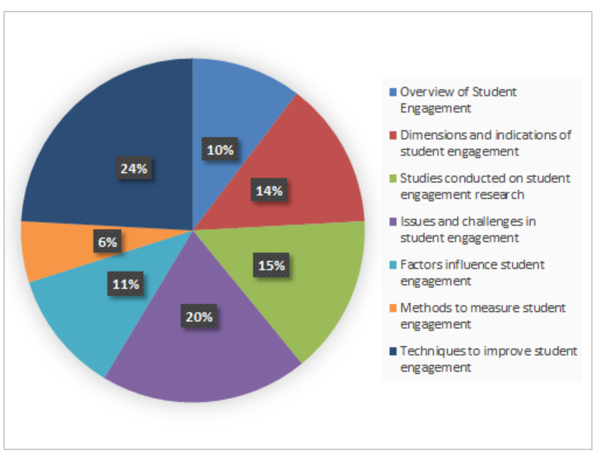


Fig. 2. Rate of the Reviewed Articles in different Categories.

Fig. 3 shows the identity of the source databases of the reviewed articles. Most of the selected articles are from Science Direct, Taylor and Francis, Sage, Springer, Wiley Online, IEEE Xplore, and IGI global, which constitute about 72% of the selected articles. Some other supporting articles have been selected from other databases.

Fig. 4 shows the number of articles in each category and according to the year of publication. Again, the results show that most of the recently published articles are about the category "techniques to improve student engagement", which clearly reflects that the problem of poor engagement is gaining more attention. The research in "Studies conducted on students' engagement" and "factors influence students' engagement" are also high. Generally, the figure shows that research in students' engagement is rapidly increasing.

Fig. 5 shows the taxonomy of literature on students' engagement. The first category is the "overview", in which two topics are discussed under this category, definitions, and history. The results show that the literature on students' engagement can be traced back to at least 70 years ago and it received more attention from researchers since the mid-1980s. The second category is the dimensions and indications. Four dimensions have been identified which are behavioral engagement, emotional engagement, cognitive engagement, and social engagement, each of which has a number of indications. The results show that emotional engagement received less attention while social engagement is just recently discussed in the literature and it requires more research.

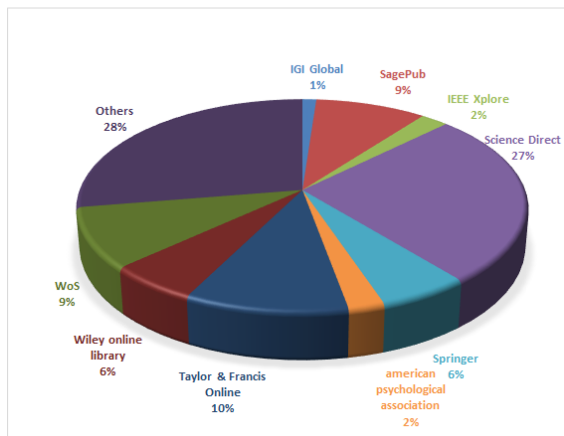


Fig. 3. Identity of the Source Databases of the Reviewed Articles.

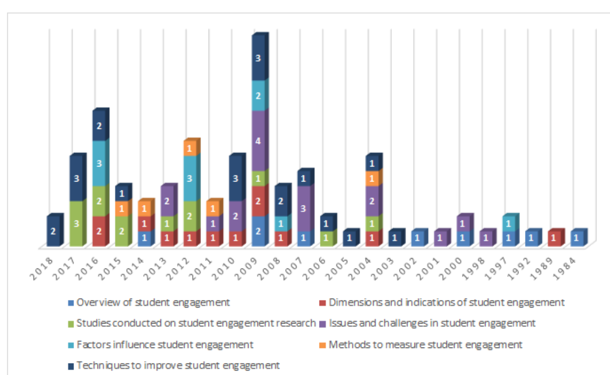


Fig. 4. Number of Articles in different Categories by Year of Publication.

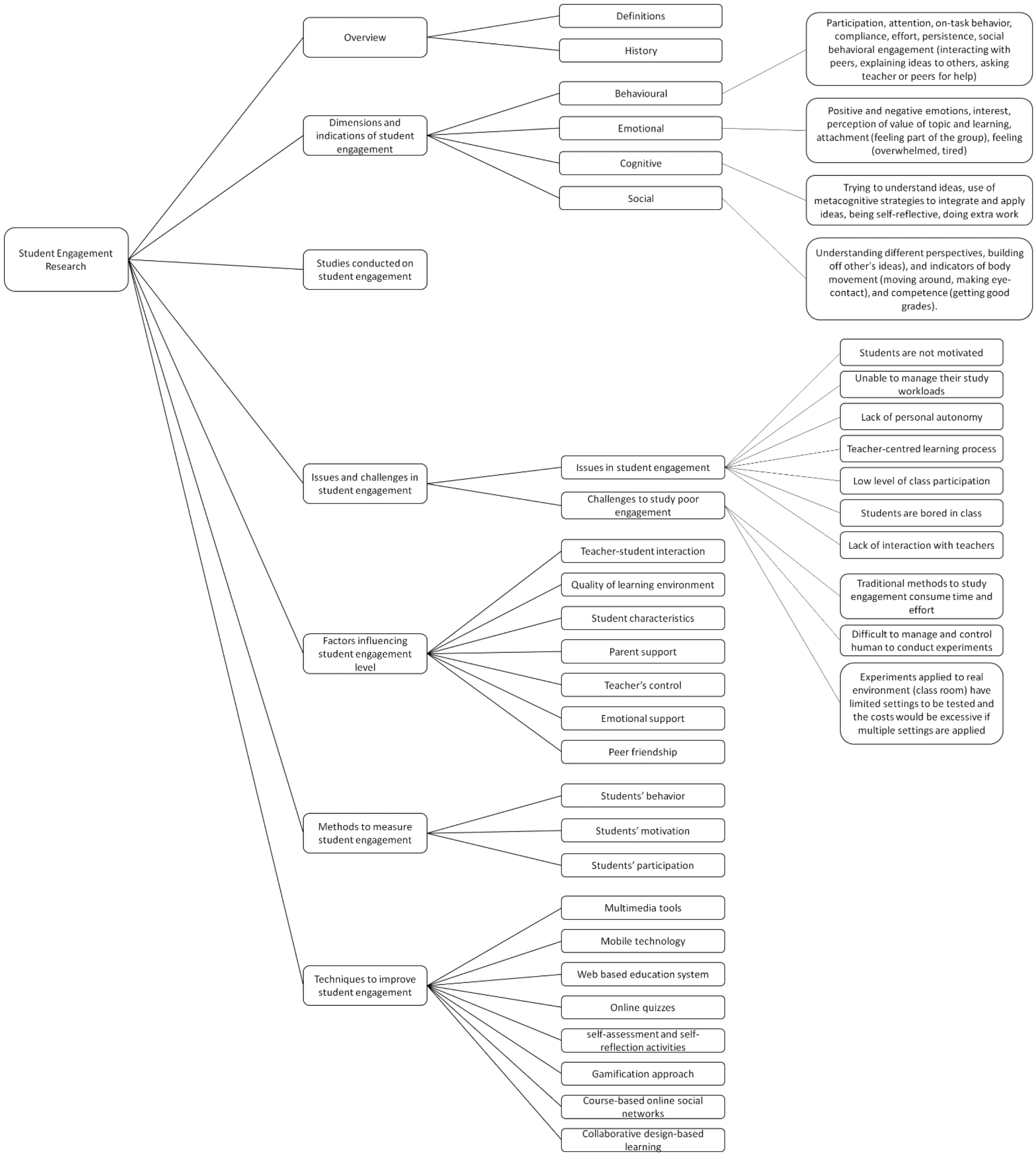


Fig. 5. Taxonomy of Literature on Students' Engagement.

The third category is "studies conducted on students' engagement". The results show that there are a number of studies that have been made either on the subject domain such as mathematics and science subjects or on the level domain such as primary or secondary school students. The results also show that there is a broad agreement in the literature that students' engagement is a multi-dimensional construct in either subject domain or level domain; the most prevalent dimensions are behavioral, emotional/ affective, and cognitive engagement. The fourth category is "Problems in students' engagement", the issues and challenges of which have been presented [26]. The results show that the issue of poor engagement often arises within middle school and aggravates during high school and becomes worse at the university level. Several issues have been revealed such as students are not motivated; unable to manage their study workloads; lack of personal autonomy; teacher-centered learning process; low-level class participation; bored in class; and lack of interactions with teachers.

The review also reveals that one of the main challenges to study these issues is the limitation of traditional research tools. Studies indicate that traditional methods consume time and effort, difficult to manage and control human subjects to conduct experiments; and experiments applied to real environment (classroom) have limited settings to be tested and the costs would be excessive if multiple settings are applied.

Under the fifth category "Factors influencing student engagement level", seven factors have been identified. The results show that teachers have a significant role among the relative factors postulated to influence students' engagement such as parents, friends, students' characteristics and learning environment. A teacher's role could be classified into a teacher's control, teacher's support, and teacher-students' interactions. The sixth category is "Methods to measure student engagement", the results of which show that there are three indicators that could be utilized to measure or evaluate students' engagement which are participation level, motivation level, and their positive or negative behaviour.

The last category is "Techniques to improve students' engagement". Several technologies have been deployed to enhance engagement and subsequently, educational outcomes. The results show positive attempts by researchers to engage students using multimedia tools, mobile technology, and web-based systems. However, these techniques might not be applicable in real situations due to the absence of one or more implementation requirements such as limitation of time, resources, teacher capability, and suitable environment.

IV. Student Engagement Research

Students who are engaged receive more teacher involvement, highly motivated to attend classes, and actively involved in classroom discussions [7]. Disengaged students are more likely to find that teachers become more controlling and increasingly withdraw their support, thereafter develop habitual truant and/or occasionally misbehave in classrooms [7].

A. Overview of Student Engagement

Kuh [8] define student engagement as "participation in educationally effective practices that lead to a range of measurable outcomes". Student engagement promotes both emotional wellbeing and learning outcomes. Research has highlighted the negative effect boredom can have on emotional wellbeing on students [9]. In Student Engagement Literature Review by Trowler [10], he states this research gained widespread attention particularly in North America and Australasia where the large scale of student engagement survey took place. Most prolific researchers such as Kuh [11] and Coates [12] developed and implemented large scale of national student engagement surveys within several universities.

According to Kuh [8], a theory of students' engagement has been deliberated in the literature for more than 70 years and attracted widespread attention of social researchers in the literature since the mid-1980s [10] [13]. Harper [14] postulated engagement "is more important than involvement or participation as it requires feelings and sense-making as well as activity". The past twenty years of research on students' engagement indicate that engagement and motivation lead to higher achievements of the students' performance [15]. In a nutshell, for a good purpose, a huge number of researchers have studied and investigated students' engagement due to deteriorating engagement over the years [16].

B. Dimensions and Indications of Student Engagement

Initially, Finn [17] introduced two-component student engagement models; Participation and Identification (PI) Model. In his terms, Participation refers to "behavioral engagement" and Identification refers to "affective/emotional engagement". Since many researchers have shown interest to pursue engagement research, the multi-dimensional model becomes more popular. Therefore, Fredricks et al. [18] suggest that three different components that construct students' engagement which are behavioral, emotional, and cognitive as shown in Fig. 6. Educational institutions has highly recognized the significance of students' engagement as it is widely noticed that many students are suffering from several disengagement indicators such as bored, unmotivated, and uninvolved from the academic and social aspects of school life [19].

The Behavioral component is reflected by student compliance with school rules such as attendance. The Emotional component concerns student's feeling that reflected by attitude, belongingness and his/her opinion towards the school. Finally, Cognitive component is reflected by learning involvement.

Recently, another engagement component has been proposed by literature named social that is reflected by social interaction in learning process [20] [21] [22].

This classification of student's engagement (behavioral, emotional, cognitive, and social) is supported by other research works such as in [23] [24] [25] they confirmed these types via interview session, it is also observed that these types have long-term impact on better academic results [27].

C. Studies Conducted on Students' Engagement Research

Previous studies have shown that levels of engagement are directly related to academic achievement in some particular subjects [17] [18] [28]. Therefore, there are few studies that aim to investigate subject-specific students' engagement in Mathematics, Science, Malaysian Studies and English Language. A study conducted by Martin et al. [29] reports that engagement in mathematics and science subjects deteriorate during the secondary school years. Hence, Wang et al. [30] develop and validate student and teacher-report survey to measure students' engagement in mathematics and science by proposing a bifactor modeling approach.

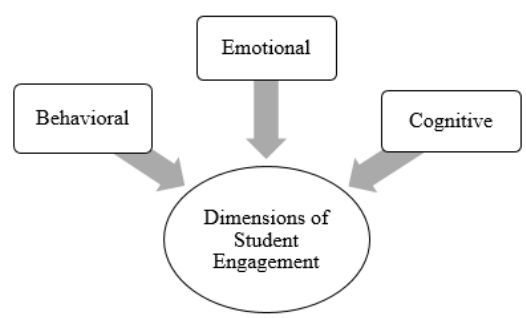


Fig. 6. Dimensions of Student Engagement.

In another study, Ayub et al. [31] investigate mathematics engagement by focusing on the cognitive, affective and behavioral engagement domains among secondary school students. A total number of 387 students from urban and rural secondary schools in Pahang, Malaysia, were randomly selected. According to PISA survey conducted in 2012, engagement performance in mathematics lesson declined among Malaysian students. Thus, in their research, they suggest that teachers should restructure their teaching strategy by creatively devising new techniques in teaching mathematics. Consequently, when planning for mathematics activities or strategies, crucial elements of engagement; cognitive, behavioral and emotional are accounted for to improve engagement in mathematics lessons. The researchers also emphasize strategies that are needed to be integrated during the teaching should not be too teacher-centered [31].

In a similar study conducted at a public university, Universiti Putra Malaysia, the researchers investigate the relationship between students' attitudes toward statistics and mathematics engagement among 293 undergraduate students who are enrolled in statistics courses from several programs [32]. Their analysis of mathematics and statistics engagement domain reveals that attitudes towards statistics are positively related to the emotional domain, then a cognitive domain and followed by a behavioural domain. Their findings show attitudes are vital for students to be engaged in mathematics and statistics [32].

In the Malaysian context, lecturers face challenges in teaching theoretical subjects like Malaysian Studies and English Language because students complain that these subjects are irrelevant and not engaging [32]. Through their study, it is found that project-based learning is an approach to improve students' engagement [33] [34]. However, a lack of commitment from students and time-consuming are the barriers to implement project-based learning. They suggest that the lecturers can alleviate the difficulties in practicing project-based learning if they are well-equipped with the knowledge and skills in executing this strategy. In another study by Willms et al. [35], two social studies teachers and 24 secondary school students are involved. From their analysis, they found that students display engagement as they engage behaviorally and emotionally and enjoy their assessment in a group.

Ishak and Amjah [36] discovered that the level of engagement is fair for the first activity as expressed by their participation and interest, and enthusiastic to work in group. The study dedicated for teachers to realize the importance of engagement in learning. Teoh et al. [37] conducted a quationnaire for a total of 64 students and they identified five indices of students' engagement "student-faculty interaction", "cooperation among students", "life-long learning", "and experience with diversity", "active learning". Further investigation reports other aspects such as imaginative teaching practice, prompt feedback to students' work, react fairly to students' performance and suitable teaching strategy would encourage involvement and commitment of students in their studies.

Even though there is a broad agreement in the literature that students' engagement is a multi-dimensional construct and some notable research use variations ranging from two to four dimension; behavioral, cognitive, emotional and social. Despite these inconsistencies, the most prevalent dimension in the literature is that engagement consists of three distinct dimensions; behavioral, emotional/affective and cognitive engagement [18]. Among these, behavioral engagement receives the most attention and emotional engagement, the least [18]. From our analysis, a few prominent researches use this crucial multi-dimensional focus in subject domains or among primary/secondary school students. However, past investigations report that tertiary level institutions face more signs of poor engagement or lack of commitment to studies [38]. Furthermore, Krauss et al., [39] report that existing studies on school engagement have been mostly limited to individual and school-based predictors in Western countries.

D. Problems in Students' Engagement

Issues of poor engagement often arise within middle school and are aggravated during high school [39]. Marks [19] and Stipek and Byler [40] report that there is a steady decline from the middle to high school students in affective interest and motivation to learn, followed by a decline in cognitive enthusiasm to face challenges in tasks. Furthermore, evidence supported by Willms et al. [41] shows that poor engagement usually becomes a concern in middle school and high school.

It was discovered by several studies conducted in Australia and the United States the performance in of undergraduates is declining as students are motivated to study and they are less involved with university activities [42]. Besides, according to Astin & Sax [43], students less requesting teacher's advice, oversleeping and absence from classes.

Another study by Thang [44] and Ming and Alias [45] conducted on public and private universities of Malaysia, the result showed that majority of students experience a teacher-centered learning process and lack personal autonomy.

In general, Asian students are suffering from low-level class of engagement and more that 80% of students do not ask questions during class this is due to their disinterest with the lessons [46] [47] [48]. Another study by UCLA discovered that 40% of students are not engaged and feel boring during class and 59% are get bored at least half of their classes [49] [50]. Another survey by Indiana University revealed that due to the lack of interaction with teachers, 30% of the students are engaged and 75% of the students are not interested with the given subjects [51] [52]. On the other hand, Researchers have used traditional methods such as survey and questionnaire to study and investigate students' engagement consume a lot of time and efforts and limited settings as repeating an experiment with different settings, gathering a huge crowd of students would be tedious [53] [54] [55].

E. Factors Influencing Students' Engagement Level

Apparently, high-level engagements in classrooms are related to enhanced achievement, attained knowledge, skill and effective learning [56]. Better engagement produces better emotional functioning [57]. The purpose of engagement in the classroom is to serve as a protector against student drop out issues and to prevent students' involvement in unhealthy activities [57]. Research findings indicate that students' engagement is correlated to academic performance, and their disengagement leads to poor academic performance in a variety of subjects or domains [12].

Recent results reveal that parent support for basic psychological needs seems to be the most significant factor in either academic motivation or dropping out cases [58]. In addition, parental monitoring and family cohesion are reported as predictors of engagement [39]. Further findings indicate that supportive relationships with lecturers are significantly associated with students' engagement [39]. According to Abdullah et al. [59], students' personality, environment, and influence of instructors and peers are factors that are able to motivate students' engagement throughout the lesson. Ricard and Pelletier [58] also agree that peer friendship affects motivation, persistence, and engagement in a classroom.

In another research, Shernoff et al. [60] use the Experience Sampling Method (ESM) to predict students' engagement and experience. Their demonstration shows that the quality of the learning environment contributes to students' engagement and a sense of classroom self-esteem. Even though students' characteristics substantively influence their motivation, learning, and quality of learning environment directly or indirectly influence the students' academic performance [61]. It is also believed that students' engagement in the classroom is influenced by teachers' control [61]. Teachers' role is essential for the quality of learning process and students' experiences [62]. In addition, teachers' support and caring have been found to be pivotal in students' engagement. Carroll et al. [63], state that the teacher is the strongest factor after students' characteristics. The teacher influences the students via a variety of activities such as assisting students, providing instructional, emotional and organizational support. Based on the findings by Lam et al. [64], among the relative factors postulated to influence students' engagement, teacher-student interaction has received significant endorsement. Fig. 7 depicts the factors that influence students' engagement.

F. *Methods to Measure Students' Engagement*

Another approach to defining student engagement is to see how it is measured. For centuries, few common measures have been practiced to identify if students are actively engaged or involved in the learning process. These measures mainly focused on the traditionally "quantifiable" aspects of student behavior, attendance rates, truancy, time-on-task, and consequently suspension/discipline rates [65]. Generally, highly motivated student always attends the classes and score standardized test. However, demotivated students always absence to class and possess poor academic performance. Researchers stressed that more systematic approach to the measurement of engagement is the most imperative for future research direction [4] [66].

Other indicators that have been used to measure engagement in the literature include teacher ratings, interviews, observations, cross-cultural data and assessments grades [6] [4] [5]. On the other hand, extra guidance and culture of explanation instead of the culture of right answer are other ways to measure students' engagement. Similarly, as reported by Fredricks & McColskey [4], students are encouraged to participate in a classroom by honestly self-reflecting their understanding of the lesson before moving onto the next lesson. For example, each student can rate their understanding from 1-3 (1 is lowest understanding rate and 3 is highest understanding rate). By rating their understanding, they can reflect what was taught.

An impressive way to measure students' engagement is to ask questions in a classroom. Typically, when a question is asked, the same person will raise each time. This issue might lead to inattention in the classroom. Thus, overall students' participation is required to measure the students' engagement [10].

G. *Computational Methods/Techniques to Improve Students' Engagement*

When teachers lack the resources, ideas, and materials to effectively manage the social and emotional challenges within the particular context of a classroom, students show poor interest in on-task behavior and performance [67]. Therefore, many researchers proposed methods or techniques to engage students with lessons despite the factors influencing it. Table 1 shows examples of techniques in the literature.

Windham [68] recommends the concept of "interaction, exploration, relevancy, multimedia and instruction" for students' engagement activities that strengthen the interaction of lecturers with their students and at the same time monitor how their students learn. Students are given the opportunity to interact with someone successful in their field as opposed to being theoretical and text-based. Applying real-life scenarios focused on keeping young students engaged and motivated. Engagement is a major condition for learning, in fact, outdated educational systems fail to sustain the interest and objectives of millennials generation which is infused with digital technologies [68]. Innovative and fun context of teaching strategy is essential to promote engagement. A variety of multimedia tools and incorporating technology into learning activities have been shown to increase students' interest, higher achievement, improved quality of work, higher students' motivation, and improved information literacy and critical thinking skills [70]. For example, using mobile devices for research and social-based learning activities have been shown to upsurge their engagement by as much as 78% [71].

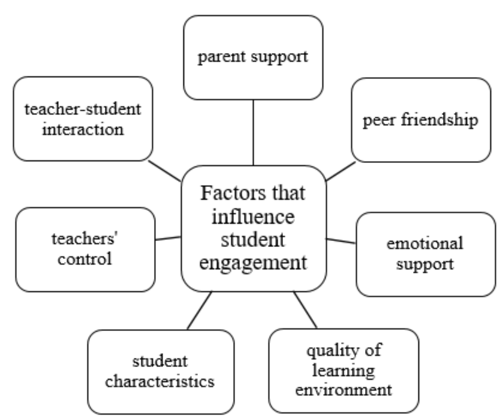


Fig. 7. Factors that Influence Student Engagement.

TABLE. I. COMPUTATIONAL METHODS / TECHNIQUES TO ENGAGE STUDENTS

Reference	Techniques
Krause & Coates [82]	▪ Stimulate discussion and debate, exploration and discovery. ▪ Stimulate question answer sessions and a range of interactive activities. ▪ Provide feedback and continuous assessment tasks early and often. ▪ Engage students in self-assessment and peer assessment. ▪ Engage students socially by asking for their opinions and acting on them. ▪ Engage students emotionally by providing examples of career success.
Tizzard [83]	▪ Use of free software such as Xtranormal to create videos as learning objects, weekly reminders, and FAQ videos. ▪ Interactive chat room experience. ▪ On-line quizzes as self-assessment tools.
Parsons & Taylor [65]	▪ Use technology in the classroom. ▪ Reduce lectures, increase group-discussion time, and ignore the “multitasking” student.
Windham 68]	▪ Interaction – Connect with experts and expertise ▪ Exploration – hands-on, inquiry-based, problem-based, and exploratory, new forms of digital media (video games, social media, and websites) ▪ Relevancy – Apply real-life scenarios ▪ Multimedia – multimedia tool (e.g. WebQuests, blogs, wikis, YouTube, video documentaries), iPhones and other mobile devices for research. ▪ Cameras, video, and video editing, projectors, sound recording equipment, animation and gaming software, and the ubiquitous PowerPoint ▪ Instruction - formative assessment
Frondeville [84]	▪ Use Quickwrites - short journal-writing assignment, summarize and predict an exam or quiz question. ▪ Use signaling to allow everyone to answer a question - asking questions that allow for multiple answers or explanations ▪ Introduce a presentation by having students pair up, talk to each other about their prior knowledge of the presentation, and generate a list of four questions.
Maheady & Gard [85]	▪ ClassWide Peer Tutoring (CWPT) is a low-cost, efficient instructional intervention that enhances student achievement by increasing opportunities for active responding and immediate feedback.
Rathvon [86]	▪ Response card instruction - students respond to teacher questions by writing on cards
Falcão et al. [69]	▪ Fun and interactive context of strategy
Heflin et al. [72]	▪ Use of mobile technologies for learning ▪ Promote collaborative learning
Goyal & Krishnamurthy [74]	▪ Use of web-based system, for example, online learning material ▪ Better organize content ▪ Cater different type of students
Russell et al. [75]	▪ Online quizzes ▪ Relevant to real-world applications
Wang [76]	▪ Self-assessment and self-reflection activities ▪ Problem-solving activities
Dim et al. [77]	▪ Gamification approach for example gamified online discussion tool
Imlawi et al. [78]	▪ Create online course-based social networks to interconnect with their students
Chen and Chiu [80]	▪ Intergroup competition mechanism and integrate it into a multi-touch platform for collaborative design-based learning (DBL).

Heflin et al. [72] investigate the impact of mobile technology on students’ attitudes, engagement, and learning. Their findings indicate that mobile technology is allied with positive students’ perceptions of collaborative learning. This clearly shows that multimedia and technology have been proven to be helpful in engaging students, exploring ways to present their ideas [71] [73].

Recently, web-based education systems have been a popular topic. Goyal and Krishnamurthy [74] suggest that cognitive strategies have a vital role to design web-based systems. Students utilize strategies when learning any subject from online learning materials. The purpose of a web-based education system is to better organize the content that caters to different characteristics of students in the learning process.

In a study by Wang [76], she identifies that a course design improves students’ behavior, emotional and cognitive engagements that promote high achievement in their studies. In addition, her findings indicate that engagement in self-assessment and self-reflection activities have an important impact on online studies and social interaction, which produce a significant engagement in problem-solving activities [76].

Ding et al. [77] emphasize on gamification approach to foster students' engagement. They developed a gamified online discussion tool, gEchoLu to examine students' engagement in online discussions. Their results indicate that gEchoLu has positive effects on student behavioral, emotional, and cognitive engagement [77].

In another study, Imlawi et al. [78] analyze the impact of students' engagement in course-based social networks and observe educational outcomes. Their findings recommend that teachers who create online course-based social networks to interconnect with their students are able to improve their engagement, motivation, and satisfaction. In addition, they provide strategies for teachers to fit their activities and enhance their students' engagement and educational outcomes.

In addition, Up [79] indicates that technology improves students' engagement including cognitive, affective, behavioral, academic, and social engagement. Therefore, lecturers need to find techniques to motivate their students to learn and increase participation rates so that their time is devoted to learning and there is no room for misbehavior. Despite the advantages of digital technology in promoting students' engagement, Chen and Chiu [80] propose and develop an intergroup competition mechanism and integrate it into a multi-touch platform for collaborative design-based learning (DBL). Their analyses show that students who participate in intergroup competition significantly foster better students' engagement, learning achievement, and creativity [81].

V. Discussion and Recommendations

From our study and analysis of the literature, we have identified several issues worth investigating.

A. *Examining the Impact of Emotion on Students' Engagement*

Despite the vast amount of research in students' engagement areas, students' emotional engagement still lacks conceptual clarity due to little attention from social studies' researchers [9] [57]. Studies on emotions have not been deeply investigated by researchers as a major factor in shaping students' engagement [9]. A few studies have examined academic emotions as a predictor for dropout, in addition to their influence on achievement [57] [87] [88]. It is recommended that investigation should be made to enhance emotional engagement in a classroom by examining a lecturer emotional states [89].

B. *Development of a Configuration Framework to Generate an Engagement-Strategy for a Particular Classroom Environment*

As discussed in Section 4, there are a number of interesting techniques and strategies that have been proposed by researchers. However, we do not expect that teachers will review all these techniques and strategies to identify the suitable and applicable to their classroom environment. Therefore, we suggest here to develop a configuration framework that is able to generate a suitable strategy from some given inputs by a teacher such as a classroom environment, subject nature, available resources, teacher's capability and so forth. To exploit all the proposed techniques and strategies efficiently, they should be compiled in a system and by using the configuration framework, the system can offer some recommendations.

C. *Utilizing the Concept of Computer-based Simulation and Virtual Environment to Study and Investigate Poor Students' Engagement Instead of using Traditional Research Tools*

Modeling and simulation gain more attention [90] [91] [92] [93], the demand on reducing time and resource costs associated with development and validation of simulation models have also increased [94] [95] [96]. Simulation is useful when the phenomenon to be studied is not directly accessible or is difficult to observe directly ([97] [98] [99]. The idea of experimenting with models rather on the real system is not something new [100] [101] [102].

One of the potential methods that could be utilized is an agent-based social simulation (ABSS) [102] [103]. ABSS is a method to model systems that comprise of individual, autonomous, cooperating agents [53] [104]. This method can be used to model human behaviors and their effects on others [105]. Social scientists have begun to convert social theories to computer programs [106]. It is then possible to simulate social processes and carry out "experiments" that would otherwise be impossible [107] [108] [109].

VI. Conclusion, Limitation and Future Works

In this paper, existing research in various categories of students' engagement is thoroughly reviewed. A collection of selected papers is classified into seven categories, overview; dimensions and indications; research studies; problems; factors influence engagement; methods to measure engagement; and techniques to improve engagement. We first describe the method to conduct this review, which includes search strategy, search terms, and selection process.

We then review the papers in each category, present the definitions and history of students' engagement under the overview category. Under the dimensions and indications category, we present the behavioural; emotional; cognitive and social dimensions and their indications. We then review 13 research studies on students' engagement, identifying issues and challenges including disengaged issues in middle school, high school, and university levels. We found that the main challenge of studying the disengagement problem is that the traditional research tools lack flexibility and the technique is tedious and costly. Subsequently, we conduct a comprehensive analysis that includes data analysis on students' engagement studies, number of selected articles by year of publication, rate of reviewed articles in different categories, number of articles in different categories by year of publication, and finally taxonomy of literature on students' engagement. This is followed by a discussion and recommendations section. In which three potential research problems are discussed. The problems are examining the impact of emotion on students' engagement; development of a configuration framework to generate an engagement-strategy for a particular classroom environment; utilizing the concept of computer-based simulation and virtual environment to study and investigate poor engagement instead of using traditional research tools. The main limitation of this review is the number of reviewed papers in some categories, albeit the papers are reliable and are broadly representative collections. Another limitation is the rapid progress and development in this field has potential to limit the timeliness of the survey.

In our future work, we shall study and analyze the importance and challenges of emotions to students' engagement to highlight factors influencing lecturers' emotions and students' emotional states that affect students' engagement in a classroom. In addition, instead of using traditional methods in studying students' engagement such as a questionnaire survey, we shall develop a simulation program that could potentially help researchers in the field of social studies to apply various environmental settings and study the impact of different strategies.

References

[1] Chapman, C., Laird, J., Ifill, N., & KewalRamani, A. (2011). Trends in High School Dropout and Completion Rates in the United States: 1972-2009. Compendium Report. NCES 2012-006. National Center for Education Statistics.

[2] Fredricks, J. A., Filsecker, M., & Lawson, M. A. (2016). Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues, Learning and Instruction, volume 23, pp 1-4.

[3] 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.

[4] 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.

[5] Castro, S., Granlund, M., & Almqvist, L. (2015). The relationship between classroom quality-related variables and engagement levels in Swedish preschool classrooms: a longitudinal study. European Early Childhood Education Research Journal, 1-14.

[6] Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. Motivation and emotion, 28(2), 147-169.

[7] Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behaviour and student engagement across the school year. Journal of educational psychology, 85(4), 571.

[8] Kuh, G. D., (2009). 'Afterword', in Harper, S. R., & Quaye, S. J., (Eds.), Student Engagement in Higher Education (New York and London, Routledge) pp. 313-18.

[9] 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.

[10] Trowler, V. (2010). Student engagement literature review. The Higher Education Academy, 11, 1-15.

[11] Kuh, George D. "What student affairs professionals need to know about student engagement." Journal of College Student Development 50, no. 6 (2009b): 683-706.

[12] Coates, H. (2007). A model of online and general campus- based student engagement. Assessment & Evaluation in Higher Education, 32(2), 121-141.

[13] Astin, A. W. (1984). Student involvement: A developmental theory for higher education. Journal of college student personnel, 25(4), 297-308.

[14] Harper, S. R. (2009). Institutional seriousness concerning Black male student engagement: Necessary conditions and collaborative partnerships. Student engagement in higher education: Theoretical perspectives and practical approaches for diverse populations, 137-156.

[15] Brewster, C., & Fager, J. (2000). Increasing student engagement and motivation: From time-on-task to homework. Portland, OR: Northwest Regional Educational Laboratory.

[16] Schlechty, P. C. (2002). Working on the Work: An Action Plan for Teachers, Principals, and Superintendents. The Jossey-Bass Education Series. Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741.

[17] Finn, J. D. (1989). Withdrawing from school. Review of educational research, 59(2), 117-142.

[18] Fredricks, J .A ., Blumenfeld, P .C . and Paris, A .H . (2004) School Engagement: Potential of the Concept, State of the Evidence . Review of Educational Research . 74 (1), pp .59–109.

[19] Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. American educational research journal, 37(1), 153-184.

[20] Fredricks, J. A., Wang, M. T., Linn, J. S., Hofkens, T. L., Sung, H., Parr, A., & Allerton, J. (2016). Using qualitative methods to develop a survey measure of math and science engagement. Learning and Instruction, 43, 5-15.

[21] Linnenbrink-Garcia, L., Rogat, T., & Koskey, K. (2011). Affect and engagement durings mall group instruction. Contemporary Educational Psychology, 36, 13-24

[22] Finn, J. D., & Zimmer, K. (2012). Student engagement: what is it and why does it matter? In S. Christenson, A. L. Reschy, & C. Wylie (Eds.), Handbook of research on student engagement (pp. 97e131). New York: Springer.

[23] Boekaerts, M. (2016). Engagement as an inherent aspect of the learning process. Learning and Instruction, 43, 76-83.

[24] Luo, W., Hughes, J. N., Liew, J., & Kwok, O. (2009). Classifying academically at-risk first graders into engagement types: Association with long-term achievement trajectories. The Elementary school journal, 109(4), 380-405.

[25] Eccles, J. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. Educational Psychologist, 44(2), 78-89.

[26] Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. Psychology in the Schools, 45(5), 369-386.

[27] Kahu, E. (2014). Increasing the emotional engagement of first year mature-aged distance students: Interest and belonging. The International Journal of the First Year in Higher Education, 5(2), 45.

[28] Akey, T. M. (2006). School Context, Student Attitudes and Behaviour, and Academic Achievement: An Exploratory Analysis. MDRC.

[29] Martin, J. Way, J. Bobis, J. Anderson. (2015). Exploring the ups and downs of math engagement in the middle school years, Journal of Early Adolescence, 35,pp. 199–244.

[30] Wang, M. T., Fredricks, J. A., Ye, F., Hofkens, T. L., & Linn, J. S. (2016). The Math and Science Engagement Scales: Scale development, validation, and psychometric properties. Learning and Instruction, 43, 16-26.

[31] Ayub, A. F. M., Yunus, A. S. M., Mahmud, R., Salim, N. R., & Sulaiman, T. (2017, January). Differences in students' mathematics engagement between gender and between rural and urban schools. In A. Kilicman, H. M. Srivastava, M. Mursaleen, & C. M. Khalique (Eds.), AIP Conference Proceedings (Vol. 1795, No. 1, p. 020025). AIP Publishing.

[32] Salim, N. R., & Ayub, A. F. M. (2017, January). Relationship between mathematics statistics engagement and attitudes towards statistics among undergraduate students in Malaysia. In A. Kilicman, H. M. Srivastava, M. Mursaleen, & C. M. Khalique (Eds.), AIP Conference Proceedings (Vol. 1795, No. 1, p. 020026). AIP Publishing.

[33] Aiedah, A. K., & Audrey, L. K. C. (2012). Application of project-based learning in students' engagement in Malaysian Studies and English Language. Journal of Interdisciplinary Research in Education (JIRE), 2(1), 37-46.

[34] Learning Point Associates (2004a). Engaged in the Learning Process, Illinois: North Central Regional Educational Laboratory (NCREL). Retrieved October 3, 2016 from http:/ /www.ncrel.org/sdrs/areas/issues/ content/cntareas/science/sc4engag.htm.

[35] Willms, J. D., Friesen, S., & Milton, P. (2009). What did you do in school today? Transforming classrooms through social, academic and intellectual engagement. (First National Report) Toronto: Canadian Education Association.

[36] Ishak, Z., & Amjah, D. Y. P. H. (2015). An Exploratory Study on Students' Engagement in Social Studies of Year 7. Journal of Management Research, 7(2), 433.

[37] Teoh, H. C., Abdullah, M. C., Roslan, S., & Daud, S. (2013). An investigation of students' engagement in a Malaysian Public University. Procedia-Social and Behavioural Sciences, 90, 142-151.

[38] Kazmi, A. (2010). Sleepwalking through Undergrad: Using Student Engagement as an Institutional Alarm Clock. College Quarterly, 13(1), n1.

[39] Krauss, S. E., Kornbluh, M., & Zeldin, S. (2017). Community predictors of school engagement: The role of families and youth-adult partnership in Malaysia. Children and Youth Services Review.

[40] Stipek, D., & Byler, P. (2004). The early childhood classroom observation measure. Early Childhood Research Quarterly, 19(3), 375-397.

[41] Willms, J. D., Friesen, S., & Milton, P. (2009). What did you do in school today? Transforming classrooms through social, academic and intellectual engagement. (First National Report) Toronto: Canadian Education Association.

[42] McInnis, C. (2001). Signs of Disengagement? The Changing Undergraduate Experience in Australian Universities. Inaugural Professorial Lecture.

[43] Astin, A. W., & Sax, L. J. (1998). How undergraduates are affected by service participation. Service Participation, 39(3), 251.

[44] 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.

[45] Ming, T. S., & Alias, A. (2007). Investigating readiness for autonomy: A comparison of Malaysian ESL undergraduates of three public universities. Journal of Reflections on English Language Teaching, 6(1), 1-18.

[46] Dasari, B. (2009). Hong Kong students' approaches to learning: cross-cultural comparisons. US-China Education Review, 6(12), 46-58.

[47] Sayadi, Z. A. (2007). An investigation into first year engineering students' oral classroom participation: a case study. Universiti Teknologi Malaysia, Faculty of Education.

[48] Mustapha, S. M., & Rahman, N. S. N. A. (2011). Classroom participation patterns: A case study of Malaysian undergraduate students. EDUCARE, 3(2).

[49] 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.

[50] 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.

[51] Yazzie-Mintz, E. (2010). Charting the path from engagement to achievement: A report on the 2009 High School Survey of Students' engagement . Bloomington, IN: Center for Evaluation & Education Policy.

[52] Nikish. C, (2013) Hobsons (http://www.hobsons.com)Students' Engagement (Http://Www.Naviance.Com/Blog/C/Stude Ntengagement).

[53] Macal, C. M., & North, M. J. (2009, December). Agent-based modeling and simulation. In Winter simulation conference (pp. 86-98). Winter Simulation Conference.

[54] Bousquet, F., & Le Page, C. (2004). Multi-agent simulations and ecosystem management: a review. Ecological modelling, 176(3), 313-332.

[55] Lawson, J., & Silver, H. (2013). A social history of education in England. Routledge.

[56] Barkatsas, A. T., Kasimatis, K., & Gialamas, V. (2009). Learning secondary mathematics with technology: Exploring the complex interrelationship between students' attitudes, engagement, gender and achievement. Computers & Education, 52(3), 562-570.

[57] Skinner, E., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic?. Journal of Educational Psychology, 100(4), 765.

[58] Ricard, N. C., & Pelletier, L. G. (2016). Dropping out of high school: The role of parent and teacher self-determination support, reciprocal friendships and academic motivation. Contemporary Educational Psychology.

[59] Abdullah, M. Y., Bakar, N. R. A., & Mahbob, M. H. (2012). Student's Participation in Classroom: What Motivates them to Speak Up? Procedia-Social and Behavioural Sciences, 51, 516-522.

[60] Shernoff, D. J., Kelly, S., Tonks, S. M., Anderson, B., Cavanagh, R. F., Sinha, S., & Abdi, B. (2016). Student engagement as a function of environmental complexity in high school classrooms. Learning and Instruction.

[61] Wubbels, T., Brekelmans, M., Mainhard, T., den Brok, P., & van Tartwijk, J. (2016). Teacher–Student Relationships and Student Achievement. Handbook of Social Influences in School Contexts: Social-Emotional, Motivation, and Cognitive Outcomes, 127.

[62] Christenson, S. L., Reschly, A. L., & Wylie, C. (2012). Handbook of research on student engagement. New York, NY: Springer].

[63] Carroll, A., Houghton, S., Wood, R., Unsworth, K., Hattie, J., Gordon, L., & Bower, J. (2009). Self-efficacy and academic achievement in Australian high school students: The mediating effects of academic aspirations and delinquency. Journal of Adolescence, 32(4), 797-817.

[64] Lam, S. F., Jimerson, S., Kikas, E., Cefai, C., Veiga, F. H., Nelson, B., & Farrell, P. (2012). Do girls and boys perceive themselves as equally engaged in school? The results of an international study from 12 countries. Journal of school psychology, 50(1), 77-94.

[65] Parsons, J., & Taylor, L. (2011). Improving student engagement. Current issues in education, 14(1).

[66] Veiga, F., Reeve, J., Wentzel, K., & Robu, V. (2014). Assessing students' engagement: A review of instruments with psychometric qualities. In I Congresso Internacional Envolvimento dos Alunos na Escola: Perspetivas da Psicologia e Educação (pp. 38-57).

[67] Marzano, R. J., Marzano, J. S., & Pickering, D. (2003). Classroom management that works: Research-based strategies for every teacher. ASCD.

[68] Windham, c. (2005). The student's Perspective. In D. Oblinger& J.Oblinger (Eds), educating the Net Generation (pp.5.1-5.16). Boulder, CO: EDUCAUSE.

[69] Falcão, T. P., Flávia Mendes De Andrade E Peres, Morais, D. C., & Oliveira, G. D. (2018). Participatory methodologies to promote student engagement in the development of educational digital games. Computers & Education, 116, 161-175. doi:10.1016/j.compedu.2017.09.006.

[70] Taylor, L., Servage, L., McRae, P., & Parsons, J. (2006). Alberta initiative for school improvement (AISI) collaborative professional development research review: Relationships, leadership and ownership. Edmonton, Alberta: University of Alberta & Alberta Education. Retrieved June, 22, 2009.

[71] Dunleavy, J. & Milton, P. (2009). What did you do in school today? Exploring the concept of Student Engagement and its implications for Teaching and Learning in Canada. Toronto: Canadian Education Association (CEA), 1-22.

[72] Heflin, H., Shewmaker, J., & Nguyen, J. (2017). Impact of mobile technology on student attitudes, engagement, and learning. Computers & Education, 107, 91-99. doi:10.1016/j.compedu.2017.01.006.

[73] Kvavik, R. B., Caruso, J. B., & Morgan, G. (2004). ECAR Study of Students and Information Technology 2004: Convenience. Connection, and Control, 5.

[74] Goyal, M., & Krishnamurthy, R. (2018). Optimizing Student Engagement in Online Learning Environments: Intuitionistic Fuzzy Logic in Student Modeling. In Optimizing Student Engagement in Online Learning Environments (pp. 187-219). IGI Global.

[75] Russell, J. E., Van Horne, S., Ward, A. S., III, E. A. B., Sipola, M., Colombo, M., & Rocheford, M. K. (2016). Large Lecture Transformation: Adopting Evidence-Based Practices to Increase Student

Engagement and Performance in an Introductory Science Course. Journal of Geoscience Education, 64(1), 37-51.

[76] Wang, F. H. (2017). An exploration of online behaviour engagement and achievement in flipped classroom supported by learning management system. Computers & Education, 114, 79-91. doi:10.1016/j.compedu.2017.06.012.

[77] Dim, L., Kim, C., & Orey, M. (2017). Studies of student engagement in gamified online discussions. Computers & Education, 115, 126-142. doi:10.1016/j.compedu.2017.06.016.

[78] Imlawi, J., Gregg, D., & Karimi, J. (2015). Student engagement in course-based social networks: The impact of instructor credibility and use of communication. Computers & Education, 88, 84-96. doi:10.1016/j.compedu.2015.04.015.

[79] Up, S. (2009). Unleashing the future: educators 'speak up'about the use of emerging technologies for learning. Recuperado el, 24.

[80] Chen, C., & Chiu, C. (2016). Employing intergroup competition in multitouch design-based learning to foster student engagement, learning achievement, and creativity. Computers & Education, 103, 99-113. doi:10.1016/j.compedu.2016.09.007.

[81] Zepke, N., & Leach, L. (2010). Improving student engagement: Ten proposals for action. Active learning in higher education, 11(3), 167-177.

[82] Krause, K. L., & Coates, H. (2008). Students' engagement in first- year university. Assessment & Evaluation in Higher Education, 33(5), 493-505.

[83] Tizzard, R. (2010, March). Strategies That Increase Student Engagement. In Society for Information Technology & Teacher Education International Conference (Vol. 2010, No. 1, pp. 934-935).

[84] Frondeville, T. D. (2009, August 03). How to Keep Kids Engaged in Class. Retrieved June 12, 2017, from https://www.edutopia.org/ classroom-student-participation-tips.

[85] Maheady, L., & Gard, J. (2010). Classwide peer tutoring: Practice, theory, research, and personal narrative. Intervention in School and Clinic, 46(2), 71-78.

[86] 86 Rathvon, N. (2008). Effective school interventions: Evidence-based strategies for improving student outcomes. Guilford Press.

[87] Kort, B., Reilly, R., & Picard, R. W. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. Proceedings of the IEEE International Conference on Advanced Learning Technologies, Los Alamitos: CA: IEEE Computer Society Press, 43-46.

[88] Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. Educational psychology review, 18(4), 315-341.

[89] Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). Contemporary educational psychology, 36(1), 36-48.

[90] Mazer, J. P. (2013). Associations among teacher communication behaviours, student interest, and engagement: A validity test. Communication Education, 62(1), 86-96.

[91] Mostafa, S. A., Darman, R., Khaleefah, S. H., Mustapha, A., Abdullah, N., & Hafit, H. (2018, June). A General Framework for Formulating Adjustable Autonomy of Multi-agent Systems by Fuzzy Logic. In KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications (pp. 23-33). Springer, Cham.

[92] Ahmed, M., Ahmad, M. S., & Yusoff, M. Z. M. (2010A, November). Modeling agent-based collaborative process. In International Conference on Computational Collective Intelligence (pp. 296-305). Springer, Berlin, Heidelberg.

[93] 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.

[94] 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.

[95] Mostafa, S. A., Ahmad, M. S., Ahmad, A., & Annamalai, M. (2013, December). Formulating situation awareness for multi-agent systems. In Advanced Computer Science Applications and Technologies (ACSAT), 2013 International Conference on(pp. 48-53). IEEE.

[96] 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.

[97] 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.

[98] Davidsson, P. (2002). Agent based social simulation: A computer science view. Journal of artificial societies and social simulation, 5(1).

[99] Mostafa, S. A., Ahmad, M. S., Ahmad, A., Annamalai, M., & Mustapha, A. (2014). A dynamic measurement of agent autonomy in the layered adjustable autonomy model. In Recent Developments in Computational Collective Intelligence(pp. 25-35). Springer, Cham.

[100] Ahmed, M., Ahmad, M. S., & Yusoff, M. Z. M. (2010B, June). Mitigating human-human collaboration problems using software agents. In KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications (pp. 203-212). Springer, Berlin, Heidelberg.

[101] Mostafa, S. A., Ahmad, M. S., Ahmad, A., Annamalai, M., & Gunasekaran, S. S. (2016, August). A Flexible Human-Agent Interaction model for supervised autonomous systems. In 2016 2nd International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR) (pp. 106-111). IEEE.

[102] 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.

[103] Gilbert, N. (2008). Agent-based models (No. 153). Sage.

[104] Mahmoud, M., Ahmad, M. S., & Yusoff, M. Z. M. (2016B). Development and implementation of a technique for norms-adaptable agents in open multi-agent communities. Journal of Systems Science and Complexity, 29(6), 1519-1537.

[105] Mostafa, S. A., Ahmad, M. S., & Mustapha, A. (2017). Adjustable autonomy: a systematic literature review. Artificial Intelligence Review, 1-38.

[106] 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.

[107] Mahmoud, M. A., Ahmad, M. S., Ahmad, A., Yusoff, M. Z. M., & Mustapha, A. (2012, June). A norms mining approach to norms detection in multi-agent systems. In 2012 International Conference on Computer & Information Science (ICCIS) (Vol. 1, pp. 458-463). IEEE.

[108] Mahmoud, M. A., Ahmad, M. S., & Yusoff, M. Z. M. (2016A, March). A norm assimilation approach for multi-agent systems in heterogeneous communities. In Asian Conference on Intelligent Information and Database Systems (pp. 354-363). Springer, Berlin, Heidelberg.

[109] Mahmoud, M. A., Ahmad, M. S., Yusoff, M. Z. M., & Idrus, A. (2015). Automated multi-agent negotiation framework for the construction domain. In Distributed Computing and Artificial Intelligence, 12th International Conference (pp. 203-210). Springer, Cham.

The Effectiveness of Stemming in the Stylometric Authorship Attribution in Arabic

Abdulfattah Omar[1]
College of Science and Humanities
Prince Sattam Bin Abdulaziz University, Saudi Arabia
Department of English, Faculty of Arts, Port Said University

Wafya Ibrahim Hamouda[2]
Department of Foreign Languages
Faculty of Education, Tanta University
Egypt

***Abstract*—The recent years have witnessed the development of numerous approaches to authorship attribution including statistical and linguistic methods. Stylometric authorship attribution, however, remains among the most widely used due to its accuracy and effectiveness. Nevertheless, many authorship problems remain unresolved in terms of Arabic. This can be attributed to different factors including linguistic peculiarities that are not usually considered in standard authorship systems. In the case of Arabic, the morphological features carry unique stylistic features that can be usefully used in testing authorship in controversial texts and writings. The hypothesis is that much of these morphological features are lost due to the execution of stemming. As such, this study is concerned with investigating the effectiveness of stemming in the stylometric applications to authorship attribution in Arabic. In so doing, three Arabic stemmers GOLD stemmer, Khoga stemmer, Light 10 stemmer are used. By way of illustration, a corpus of 2400 news articles written by different 97 authors is designed. To evaluate the effectiveness of stemming, the selected articles (both stemmed and unstemmed texts) are clustered using cluster analysis methods. Comparisons are made between clustering structures based on stemmed and unstemmed datasets. The results indicate that stemming has negative impacts on the accuracy of the clustering performance and thus on the reliability of stylometric authorship testing in Arabic. The peculiar stylistic features of the affixation processes in Arabic can, thus, be usefully used for improving the performance of authorship attribution applications in Arabic. It can be finally concluded that stemming is not effective in the stylometric authorship applications in Arabic.**

Keywords—Authorship attribution; cluster analysis; GOLD stemmer; Khoga stemmer; Light 10 stemmer; stemming; stylometry

I. Introduction

The recent years have witnessed an increasing use of stylometric approaches in addressing different authorship problems. These have been mainly based on the investigation of the lexical (e.g. frequency of distinctive words, discourse markers, and modal verbs) and structural (e.g. use of chunks, type of sentence, and sentence length) properties of the texts as a clue for identifying authors of controversial texts. In spite of the success of these approaches in solving different authorship problems of various controversial documents, presently, they are ineffective and thus, unreliable in addressing authorship problems in Arabic. This can be attributed to the peculiar linguistic features of Arabic. As thus, this study investigates the effectiveness of stemming in the stylometric authorship attribution in Arabic. The hypothesis is that derivational and inflectional morphemes (which are normally removed in stemming applications) carry unique stylistic features that may be useful in identifying authors of controversial texts. By way of illustration, this study is based on a corpus of 2400 news articles written by 97 authors derived from four newspapers. Three stemming algorithms were used. To evaluate the effectiveness of stemming, datasets (both stemmed and unstemmed texts) were clustered using hierarchical cluster analysis methods. The remainder of this article is organized as follows. Section 2 asks the research question concerning the effectiveness of stemming in the stylometric authorship applications in Arabic. Section 3 surveys the authorship attribution literature and the emergence of stylometric approaches in authorship studies. Section 4 defines the methods and procedures. Section 5 is an experimental analysis of the effect of stemming on the accuracy and reliability of text clustering performance. Section 6 is conclusion.

II. Research Question

Stemming can be broadly defined as the practice of conflating semantically equivalent word variants into the same root by removing derivational and inflectional affixes [1-3]. Technically speaking, it is a procedure that tries to remove inflectional and derivational suffixes to conflate word variants into the same stem or root [4]. The basic concept of stemming is that words of identical stem or root that refer to the same concept must, therefore, be grouped under the same type. Paice makes it clear that the function of stemming is to conflate all words which share equivalent semantics and share identical stems [5].

Many stemmers have been understood and through this, developed for a colossal range of languages including English, French, German, Dutch, Swedish, Latin, Malay, Indonesian, Slovene, Turkish, Arabic, and Hebrew. Leah, Lisa, et al. [6] point out that stemmers tend to be bespoke and exclusive to each independent language [6]. Building stemmers accordingly requires some linguistic knowledge of the language and an understanding of the needs of information retrieval. The concept of all stemmers is the reduction of the corpora size so that data mining processes (e.g. Information Retrieval, text clustering, etc.) systems work faster and more effectively. All stemmers have one thing in common: they are all designed to remove derivational and inflectional affixes and conflate word variants of the same base into a common term. Some stemmers are designed for both derivational and inflectional affixes; others are designed for only suffixes, and some are designed solely for handling simple plurals. In English, just like many Western European languages, stemming is predominantly a methodology of removing a suffix. That is, stemming is a procedure for removing suffixes attached at the end of words. The point is that stemming algorithms for English and other European languages normally fail to consider prefixes and infixes. In this, English stemming is primarily concerned with the morphology of suffixes.

With regard to Arabic, two main approaches have been developed. These are the root-based method and the light stemming approach. Elrajubi [7] explains that light stemming algorithms have the purpose to only remove prefixes and suffixes from the words, whereas root-based algorithms eliminate prefixes, suffixes and infixes. Due to the numerous problems caused by root-based algorithms, there is a tendency to use light stemmers. Light stemmers are more concerned with removing the prefix and suffix of a word [8]. Obvious examples are Khoga stemmer, Light 10 stemmer, and GOLD stemmer. In spite of the increasing use of stemming as a requirement or a pre-processing step in different NLP applications, there is no stemming algorithm that is 100% precise. To address this problem, dissimilar studies have been recently focussed on evaluating and comparing the performance of Arabic stemmers to provide users and researchers with answers about the most appropriate algorithm for their tasks [7, 9, 10]. Nevertheless, there are no definite answers to the effectiveness of stemming in stylometric authorship applications in Arabic. In the face of this problem, this study asks this research question: What is the effectiveness of stemming in stylometric authorship applications in Arabic?

III. Literature Review

Authorship attribution, also known as authorship recognition, is the process of looking for salient features in a piece of writing that relates the work to its author. Craig [11] points out that authorship studies have objectives of 'yes or no' declarations to present problems, and are said to avoid observable features if possible; due to operating at the base strata of language, where imitation or deliberate variation can be rejected [11]. The idea of authorship attribution is very old. Love [12] mentions that it ventures back to the period of the well-renowned library of Alexandria and accordingly, comprise the construction of the "Jewish and Christian biblical canons". The motive behind authorship attribution studies is that many works were written anonymously, and many others raise suspicion about their real author, and ultimately, historical evidence is sparse or indeed lacking. Traditionally, work on authorship attribution was conceived as an organized scholarly enterprise where it was not the achievements of an expert or scholar in authorship, but the contributions of a scholar to which the fortitude of authorship had constantly been a vital constituent into other investigative natures [12]. There are many examples where the task of identifying the author of a particular document was the job of politicians, journalists, and lawyers [13, 14]. Studies in this tradition often used criteria for relating works to authors on chronological and epistemological bases. One problem with such methods is that it [12] is often difficult to find reliable historical facts or knowledge-based evidence that will help in the identification of authors. Furthermore, these studies were based on what can be considered philological approaches making no use of replicable methods and thus the results were not objective and thus unreliable.

In the face of these limitations, empirically-driven approaches for authorship attribution problems were developed. The claim was that authorship attribution applications should be algorithmically processed without any reference to existing analytical results or personal knowledge of authors [15]. The mainstream of these approaches is described in the literature as stylometry. Stylometry is a quantitative inquiry into the individualities of an author's style and technique. Laan [16] defines the term as a technique with a purpose to maintain the comprehension of the, often elusive, characteristics and individualities of an author's style, or at least a fraction of it, through enumerating some of its features and qualities [16]. Merriam and Matthews [17] claiming "stylometry attempts to capture quantitatively the essence of an individual's use of language" [17]. Stylometric studies have been mainly based on computational and quantitative methods in order to reach solid conclusions regarding the authorship of a given text [18]. Accordingly, numerous studies have come to give empirical solutions to different controversial authorship issues using quantitative methods for investigating the stylistic and linguistic properties of authors.

One of the pioneering examples of the use of stylometric analysis in authorship problems is Mosteller and Wallace [19] attempt to give internal evidence for the authors of the dubious Federalist Papers based on linguistic and stylistic properties of the authors. These are 77 Federalist Papers written during 1787-1788 to Alexander Hamilton, John Jay and James Madison. These papers were published in newspapers under the pseudonym of Publius until they were collected with eight more articles to form a volume. There was a consensus about the authorship of these Papers that John Jay had authored 5 papers in the volume; while Hamilton authored 51 papers; Madison wrote 14 and both Madison and Hamilton co-authored three. The authorship of 12 papers in the volume was rather disputed since it was difficult to find out which of the two, Madison or Hamilton had authored those Papers [20, 21]. On their part, Mosteller and Wallace [19], employed tools of statistical analysis in order to investigate the mystery of authorship of the Federalist papers in the early 1960s, using function words as discriminators. The objectivity and replicability of the proposed approach opened the way to the computerized age of authorship attribution.

The basic assumption behind stylometric testing of authorship attribution is that, Holmes [22] contends, authors have unawareness of their characteristic styles. These are styles which cannot deliberately be influenced yet acquire features which are reckonable and are thus, highly unique [22] and the identification of such personal distinctive linguistic and stylistic features makes it possible to detect an author's signature and distinguish the writing of one author from another or others. In this way, researchers and particularly the statisticians, Knaap and Grootjen [23] had a tendency to investigate the lexical features of texts in order to make predictions about possible authors. As thus, the search for the most frequent words has been one of the most widely used methods for determining the author of a given work [24, 25]. Garcia and Martin [26] explain that statisticians attempted over the last decade to solve some controversial authorship problems by finding a formula grounded on the computation of tokens, word-types, and most frequently-used words. They contend that computational statisticians have tended to investigate, what they call, the 'Lexical Richness' of authors in order to propose a reliable approach to authorship attribution. On the other hand, Morton [27] argued that the use of rare words is a good indication for determining the author of a given text as this enables one writer to be prominent from another. He explains that occurring words communicate a multitude of essentials, which acquired the belief to demonstrate brilliance in writing. These were noted as "the range of a writer's interests, the precision of his observation and the imaginative power of his comparisons", and thus, exhibit his command of pattern and of interchanges [27]. Similarly, Blatt [28] asserts that rare words are quite noticeable and can be considered the writer's favorite words which makes it easier and accurate to use them as an indicator for determining authors.

The ineffectiveness of the lexical representation of texts in resolving different authorship problems, however, has led to the development of new methods. The lexical representation of texts has come to be known today as the traditional way of doing authorship attribution. It has been criticized for its ineffectiveness in providing solutions for the practical applications of authorship attribution [29, 30]. The claim is that isolated or single words are not enough for assigning disputed texts to their possible writers. The idea is simply that single words are not enough to capture the structure of documents. Different studies, therefore, have been more concerned with the morphological, syntactic, and structural features of texts (e.g. morphologically complex words, use of function words, sentence length, compounding, and punctuation).

In spite of the reasonable success of the stylometric methodologies in providing answers for many authorship problems, verifying the authorship of Arabic texts still represents a real challenge for the practical applications of author identification. This may be due to the fact that very few studies have been concerned with authorship attribution in Arabic, in which differences in language systems represent further challenges. This study tends to address this gap in literature by investigating the effectiveness of stemming in the stylometric authorship attribution in Arabic.

IV. Methods and Procedures

To evaluate the effectiveness of stemming in stylometric authorship attribution, three Arabic stemmers were used. These are Khoga stemmer, Light 10 stemmer and GOLD [31]. The rationale is that these stemmers are fast and straightforward algorithms and they are widely used in different NLP applications including information retrieval (IR) and document clustering. Although different studies have pointed to the idea that Light 10 outperforms many other stemmers, it was thought that it would be appropriate to include different stemmers for validity purposes.

This study is based on a corpus of 2400 news articles written by 97 authors derived from three newspapers. These are Al-Ahram, Asharq Al-Awsat, and Al-Hayat. Articles covered the period between 2016 and 2018. All selected articles are written in Modern Standard Arabic (MSA). This is an overly formal version of Arabic and differs substantially from spoken dialects. The rationale is that in MSA, core grammar and vocabulary remain constant. The root-and-pattern system is almost the same in all MSA dialects.

For assigning texts to their authors, cluster analysis methods were used. Cluster analysis is widely acknowledged as a successful technique for organizing any unorganized sets of documents [32]. It is an exploratory multivariate technique for systematically finding relatively homogeneous clusters of cases based on proximity measures without prior assumptions about differences within sets of data investigated [33-35]. It is a deterministic process that identifies discrete categories under any inherent structure in the data [36-41]. It is thus an inductive technique that explicitly attempts to group data sets into discrete classes [42, 43]. The aim of cluster analysis can be summarised as grouping a collection of objects into subsets where members of each subgroup are more closely related to one another than members assigned to the other group/s. Groups are technically referred to as clusters. Given a corpus of 2400 documents, these can be clustered where members of each cluster share specific characteristics. In authorship recognition applications, the assumption is that texts grouped together are more likely to be written by the same author. To perform cluster analysis, Euclidian distance, being a straightforward measure, was used. Euclidian distance is the most widely used and is reported to provide reliable results in general. As for the clustering method, Ward linkage is used. The rationale is that the Ward linkage clustering (or what is usually referred to as increase in sum of squares) with Euclidean measure seems to be the most convenient for the present case because it makes the clearest partitioning of the matrix rows.

V. Analysis

In order to evaluate the effectiveness of stemming in stylometric authorship applications, two processes were carried out. First, similar texts were grouped together assuming that texts grouped together are more likely to be written by the same author. Second, clustering structures were compared to the bibliographic information of each author. In order to compute the similarity between texts and group similar texts together, the Ward linkage clustering method with Euclidean distance measure was used. As a result, the matrix rows are assigned to clearly identified four groups. One advantage of this clustering is that it offers a solution for a traditional problem in cluster analysis-the decision of the optimal number of clusters that fits a dataset. The strong tendency towards left-branching that is associated with other clustering methods is avoided with Ward clustering. The matrix rows are assigned to four main groups as shown in Fig. 1.

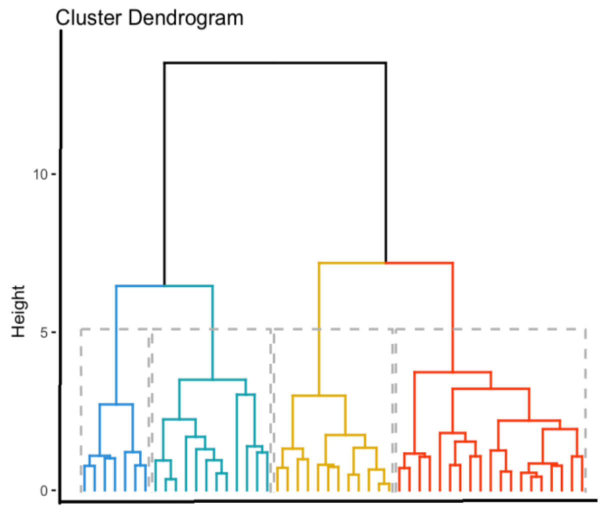


Fig. 1. Hierarchical Cluster Analysis of the Data Matrix.

For clustering validity purposes, two approaches were used. These were cross-validation and relative comparison. The purpose was to validate the foregoing analysis by seeing whether the same analytical methods applied to an alternative representation of the data give identical or at least similar results. In a cross-validation approach, the texts were randomly divided into two subsets, say *A* and *B*, and the cluster analysis is carried out separately on each of *A* and *B*. The similarity of the results is the indication of validity [44]. Comparison shows a close fit between the results as there is a total correspondence between the structures based on the data matrix composed of all the 2400 rows and the structures based on the random distribution of these 2400 rows into two groups.

For relative comparison analysis, a relative approach was based on comparing the clustering structure, generated by the same algorithms but using an alternative representation of the data; this was done by cluster analyzing a principal component reduction of the data matrix. Analysis showed that there is a close fit between the two clustering structures despite the minor differences. Consequently, it can be claimed that the agreement between the clustering structures supports the validity of hierarchical cluster analysis results. As a final step, the clustering structure obtained here was compared to the bibliographic information of each author. Members of each class or cluster were compared to the stories of each author.

The analysis was carried out four times and yielded specific results, namely, that the Light 10 dataset brought a 67% accuracy rate. The Khoga stemmer dataset yielded a 64% accuracy rate, whilst the GOLD and "Without stemming" datasets yielded accuracy rates of 61% and 78%, respectively.

The findings indicate clearly that clustering performance works much better without executing stemming. This can be attributed to two reasons. First, stemmers commit numerous errors. The three stemmers merged words that are different in form and are also semantically distinct and different from each other. Furthermore, stemmers find no solutions to homographs. This means that stemmers conflate word forms that are completely different in meaning. It was found out that the stemmers made two kinds of error: over-stemming and under-stemming. Over-stemming refers to forming larger stem classes where unrelated forms are wrongly conflated. Under-stemming, on the other hand, refers to failing to conflate variant forms of the same stem leaving them ungrouped.

It was found out that clustering structures based on Gold stemmer were the least reliable. This is due to the so many problems associated with this stemmer. These problems can be summarized as follows. First, it removes only one suffix from a word, due to its nature as a single pass algorithm. Second, it fails to form words from the stems, or matches the stems of like meaning words. Finally, its large set of rules and the recoding stage affect the speed of execution. The Khoga stemmer comes second in terms of the effectiveness. One good advantage of this stemmer attempts to find solutions to irregularities or what the compilers call non-formulaic changes (i. e. irregular plurals) by providing a lexicon within the stemmer. The problem with such non-formulaic changes is that they are unpredictable and stemming without the usage of a lexicon is fundamentally unmanageable without presenting errors.

Based on the quantitative results of the performance of each stemmer, Light 10 Stemmer can be claimed to be the most effective stemmer for Arabic data in relation to stylometric applications. This stemming algorithm is a procedure for removing the derivational and inflectional suffixes from Arabic words. However, putting the algorithm into practice and test, it is observed that it has some shortcomings. First, it makes the two kinds of errors of over-stemming and under-stemming. That is, it is sometimes too aggressive in conflation and groups words like execute and executive together and sometimes it is too weak and misses words so that they are not conflated. Second, it articulates terms (stems) that are not words and are too difficult to identify. Finally, it ignores prefix removal completely.

The second reason that can be attributed to the poor performance of the text clustering based on stemmed entries is the peculiar nature of the morphological system in Arabic. With stemming, the morphological features of Arabic, which carry unique stylistic features that distinguish authors, are lost to a great extent.

VI. Conclusion

This study addressed the question of the low performance of stylometric authorship applications of Arabic texts. The hypothesis was that part of the problem is related to the lack of consideration to linguistic peculiarities that are not usually considered in standard authorship systems. In light of this argument, the study investigated the effectiveness of stemming which is a prerequisite in different stylometric authorship systems on the performance of stylometric authorship systems in Arabic. The results indicate clearly that morphological information can be usefully used for improving the performance of authorship attribution and detection in Arabic texts due to the unique stylistic features of the affixation processes in Arabic. Controversial texts in Arabic can, thus, be assigned to their authors based on detecting stable morphological patterns with reliable authorship performance. Although the proposed system was tested only on literary texts written in Standard Arabic, the implications of the study can be practically used for the authorship problems in other text genres including emails, newsgroup messages, Facebook posts, and tweets as well as different Arabic varieties which still represent a real challenge for the practical applications of author identification.

Acknowledgment

I 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 fulfillment of the current research project.

References

[1] L. C. Jain, H. S. Behera, J. K. Mandal, and D. P. Mohapatra, Computational Intelligence in Data Mining - Volume 3: Proceedings of the International Conference on CIDM, 20-21 December 2014. Springer India, 2014.

[2] M. Bramer, Principles of Data Mining. Springer London, 2007.

[3] R. Feldman, J. Sanger, and C. U. Press, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2007.

[4] J. Savoy, "A stemming procedure and stopword list for general French corpora," J. Am. Soc. Inf. Sci., vol. 50 no. 10, pp. 944-952, 1999.

[5] C. D. Paice, "Method for evaluation of stemming algorithms based on error counting," J. Am. Soc. Inf. Sci., vol. 47, no. 8, pp. 632-649, 1996.

[6] S. L. Leah, B. Lisa, and E. C. Margaret, "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis," Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275-282, 2002.

[7] O. M. Elrajubi, "An improved Arabic light stemmer," in 2013 International Conference on Research and Innovation in Information Systems (ICRIIS), 2013, pp. 33-38.

[8] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Light Stemming for Arabic Information Retrieval," in Arabic Computational Morphology: Knowledge-based and Empirical Methods, A. Soudi, A. v. d. Bosch, and G. Neumann, Eds. Dordrecht: Springer Netherlands, 2007, pp. 221-243.

[9] Y. Jaafar, D. Namly, K. Bouzoubaa, and A. Yousfi, "Enhancing Arabic stemming process using resources and benchmarking tools," Journal of King Saud University - Computer and Information Sciences, vol. 29, no. 2, pp. 164-170, 2017/04/01/ 2017.

[10] Y. Al-Lahham, K. Matarneh, and M. Hassan, "Conditional Arabic Light Stemmer: CondLight," International Arab Journal of Information Technology, vol. 17, 12/05 2018.

[11] H. Craig, "Stylistic Analysis and Authorship Studies," in A Companion to Digital Humanities, S. Schreibman, R. Siemens, and J. Unsworth, Eds. Oxford: Blackwell, 2004, pp. 273-288.

[12] H. Love, Attributing Authorship: An Introduction. Cambridge: Cambridge University Press, 2002, p. 240p.

[13] P. Juola, J. Sofko, and P. Brennan, "A Prototype for Authorship Attribution Studies," Lit Linguist Computing, vol. 21 (2), no. 2, pp. 169-178, June 1, 2006 2006.

[14] P. Juola, "Authorship Attribution," Foundations and Trends R in Information Retrieval, vol. 1, no. 3, pp. 233-334, 2008.

[15] H. Moisl, "Using electronic corpora in historical dialectology research," in Studies in English and European Historical Dialectology, M. Dossena and R. Lass, Eds. Brussels; Frankfurt: Peter Lang, 2009, pp. 68-90.

[16] N. M. Laan, "Stylometry and Method. The Case of Euripides," Lit Linguist Computing, vol. 10, no. 4, pp. 271-278, November 1, 1995 1995.

[17] T. Merriam and R. Matthews, "Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe," Literary and Linguistic Computing, vol. 9, no. 1, pp. 1-6, 1994.

[18] G. Tambouratzis and M. Vassiliou, "Employing Thematic Variables for Enhancing Classification Accuracy Within Author Discrimination Experiments," Lit Linguist Computing, vol. 22, no. 2, pp. 207-224, June 1, 2007 2007.

[19] F. Mosteller and D. L. Wallace, Inference and Disputed Authorship: The Federalist (Addison-Wesley series in behavioral sciences. Quantitative methods.). Reading, Mass.: Addison-Wesley Pub. Co., 1964, pp. xv, 287 p.

[20] J. Rudman, "The Twelve Disputed 'Federalist' Papers: A Case for Collaboration," Proceedings Digital Humanities, pp. 353–356, 2012.

[21] J. Savoy, "The Federalist Papers revisited: A collaborative attribution scheme," in Proceedings of the Association for Information Science and Technology, Montreal, Quebec, Canada., 2013.

[22] D. Holmes, "The Evolution of Stylometry in Humanities Scholarship," Literary and Linguistic Computing vol. 13, pp. 111-117, 1998.

[23] L. v. d. Knaap and F. A. Grootjen, " Author identification in chatlogs using formal concept analysis," presented at the Proceedings of the 19th Belgian-Dutch Conference on Artificial Intelligence (BNAIC2007), Utrecht, The Netherlands, November 2007, 2007.

[24] J. F. Burrows, "Questions of Authorship: Attribution and Beyond A Lecture Delivered on the Occasion of the Roberto Busa Award ACH-ALLC 2001, New York," Computers and the Humanities, vol. 37, no. 1, pp. 5-32, 2003.

[25] J. F. Burrows, "All the Way Through: Testing for Authorship in Different Frequency Strata," Lit Linguist Computing, vol. 22, no. 1, pp. 27-47, April 1, 2007 2007.

[26] A. M. Garcia and J. C. Martin, "Function Words in Authorship Attribution Studies," Lit Linguist Computing, vol. 22, no. 1, pp. 49-66, April 1, 2007 2007.

[27] A. Q. Morton, "Once. A Test of Authorship Based on Words which are not Repeated in the Sample," Lit Linguist Computing, vol. 1, no. 1, pp. 1-8, January 1, 1986 1986.

[28] B. Blatt, Nabokov's Favorite Word Is Mauve: What the Numbers Reveal About the Classics, Bestsellers, and Our Own Writin. Simon & Schuster, 2017.

[29] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods," Journal of American Society for Information Science and Technology, vol. 60, no. 3, pp. 538-556, 2009.

[30] M. S. Tamboli and R. S. Prasad, "Authorship Analysis and Identification Techniques: A Review," International Journal of Computer Applications, vol. 77, no. 16, pp. 11-15 2013.

[31] M. Rogati, S. McCarley, and Y. Yang, "Unsupervised learning of Arabic stemming using a parallel corpus," in Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Stroudsburg, USA, 2003, vol. 1.

[32] H. Moisl, Cluster Analysis for Corpus Linguistics. Berlin, Munich, Boston: Walter de Gruyter, 2015.

[33] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, New Jersey: John Wiley& Sons, INC, 1990.

[34] A. Fielding, Cluster and Classification Techniques for the Biosciences. Cambridge, UK ; New York: Cambridge University Press, 2007, pp. xii, 246 p.

[35] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

[36] G. W. Milligan, "Clustering Validation: Results and Implications for Applied Analyses," in Classification and Clustering, P. Arabie, Hubert, L.J. and De Soete, G, Ed. River Edge, NJ: World Scientific Publishing Co Pte Ltd, 1996.

[37] B. Everitt, S. Landau, and M. Leese, Cluster analysis, 4th ed. / Brian S. Everitt, Sabine Landau, Morven Leese. ed. London: Arnold ; New York : Oxford University Press, 2001, pp. viii, 237 p.

[38] G. Punj and D. W. Stewart, "Cluster Analysis in Marketing Research: Review and Suggestions for Application," Journal of Marketing Research, vol. 20, no. 2, pp. 134-148, 1983.

[39] J. F. Hair, Multivariate data analysis, 6th ed. ed. Upper Saddle River, N.J. ;London: Prentice Hall PTR, 2006, pp. xxiv, 899 p.

[40] M. R. Anderberg, Cluster analysis for applications (Probability and mathematical statistics ; 19). New York ; London: Academic Press, 1973, pp. xiii,359p.

[41] B. Everitt, Cluster analysis, 3rd ed. London: E. Arnold, 1993, pp. viii, 170.

[42] R. Adams, "Perceptions of innovations: exploring and developing innovation classification," PhD, School of Management Cranfield University, 2003.

[43] B. Mirkin, Clustering for Data Mining: A Data Recovery Approach (Computer Science and Data Analysis Series). Taylor & Francis Group, LLC, 2005, p. 266.

[44] A. C. Rencher, Methods of Multivariate Analysis, Second Edition ed. John Wiley & Sons, INC, 2002.

Edge IoT Networking

Majed Mohaia Alhaisoni
Department of Computer Sceince
University of Ha'il, Ha'il
Saudi Arabia

***Abstract*—Data transmission has witnessed a new wave of emerging technologies such as IoT. This new way of communication could be done through smart communication such as smart sensors and actuators. Thus, data traffic keeps traversing to the main servers in order to accomplish the tasks at the sensors side. However, this way of communication has encountered certain issues related to network due to the nature of routing forth and back from the end users to the main servers. Subsequently, this incurs high delay and packet loss which successively degrades the overall Quality of Service (QoS). On the other hand, the new way of data transmission, which is called "edge IoT network", has not only helped on reducing the load over the network but also made the nodes to be more self-manage at the edge. However, this approach has some limitations due to the power consumption and efficiency, which would lead to node failure and data loss. Therefore, this paper presents a new model of combining network science and computer network in order to enhance the edge IoT efficiency. Simulation results have shown a clear evidence in improving the efficiency, communicability, degree, and overall closeness.**

Keywords—Edge IoT; network centrality; communicability; degree; closeness

I. Introduction

In the past few decades, the advances in technology is increasing very rapidly where end terminals have become well-spread over the network. The massive development in technologies, such as wireless communications and mobiles have created a platform for people to exchange information easily. Moreover, lightweight smart devices are going up, not only that but also daily life has become more interconnected across things [1]. Hence, wireless sensors network (WSN) is one of the most promising and convenient for data gathering, mainly at the era of Internet of things (IoT) [2]. However, such heterogeneous in data communication and gathering shall bring more burden over the network. Consequently, big data is more generated which needs careful consideration in terms of data classification and analysis [3].

IoT is nowadays being active technology where things are all interconnected at home, road, and buildings. Hence, actuators, sensors are commonly used to interact between devices for data transmission and processing [4]. However, since sensors are tied up with life span due to the battery lifetime and energy, this cause a critical hurdle for such type of communication to be more spread and utilized in communication and networking [5]. Therefore, various techniques were proposed in literature to address the issues related to power consumption and life span of WSN such as [6][7][8]. On the other hand, WSN has faced some issues and limitations on communicating with servers and cloud due to the sensitivity of data which is intolerable of low data rate and high latency [9]. Thus, these two key factors are very critical for IoT where massive computation and high traffic needs to be transferred timely to the end users. Henceforward, Edge IoT paradigm has become the most interesting way of communication which could help on alleviating previous highlighted issues related to communication between WSN and cloud [10]. In this approach, technology is moving from centralized cloud into distributed Edge nodes at the network. Therefore, nodes act as a client-server simultaneously which acts on receiving requests and processing at the same time with no need of interconnecting with the cloud [11]. This new way of interconnection has undoubtedly improved the latency and efficiency. However, edge devices encounter limitations of computing resources and life span at the network, which cause data loss and halted issues for data transmission, and processing.

This paper looks at the efficiency and load distribution issues at the edge IoT network. Hence, a new algorithm has been proposed to overcome these limitations. Thus, It combines network science and computer network in order to get the most use of it. Therefore, since edge devices are spread at the network edge, It is very important to consider the network centrality and load distribution. In other words, devices at the edge are not supposed to be constantly interconnected to many devices and serving them concurrently. This would easily abuse resources and make the nodes in a risk of failure at any time, furthermore, date rate would also be affected which shall degrade the QoS and response time for request by other end users. Results have shown a clear evidence in enhancing the centrality, communicability, and overall closeness.

II. Related Work

Edge computing has gained more attention recently since it is still evolving as a technology paradigm. Therefore, Nokia and IBM have launched the first edge computing at early 2013. Following that, auspices of the European Telecommunications Standards Institute (ETSI) has introduced a mobile edge computing. Moreover, in 2015 opencomputingedge.org was launched by Vodafone, Intel, Huwai in cooperation with Carnegie Mellon University (CMU) [12]. On the other hand, research has focused on studying different paradigms of edge IoT, and discussed various issues and limitation on edge IoT. In [13] edge computing has been proved that it is very useful as a local server, mainly at complex core network. Moreover, a third-party application was implemented as a local server and shown a better performance.

Another study in [14] has proposed a game theory approach to offload the computation in distributed edge IoT rather than cloud, their results have shown that IoT is more effective on computational rather cloud based. Moreover, device-to-device at edge has been proposed in [15] where they have clearly shown that offloading resources at edge enhance the efficiency. Furthermore, placing cloudlet at edge approach has been introduced by [16] where they have traded-off between the cost of placing these cloudlets and the gained end-to-end delay. In their approach, workload is assigned to the best-located cloudlet in order to reduce the response time for users' requests. Additionally, resources allocation at edge IoT is discussed in [17] where they have shown how limited resources can be optimized with consideration of plan capacity at the edge. However, existing work has analyzed the disadvantages of moving data from cloud to the end users where latency is one of the main hurdles in such approach. Therefore, in [18] authors have proposed a cloudlet to run the big data instead of adopting cloud model. This approach has shown an effective way of enhancing the communication between end users and big data applications.

Furthermore, GigaSight model has been introduced by [19] in order to minimize the latency between end users and cloudlet in processing the video while the bandwidth is reserved within the core network. In addition to previous studies, authors in [20] have proposed a mobile edge IoT where Base station is connected to a fog node which act locally by handling the data without the need to connect remotely to cloud. However, in order to alleviate the overhead on forwarding the packets among nodes, SDN based was utilized on top of fog nodes. This approach has been effective on reducing the end-to-end delay. Moreover, end -to-end delay reduction among end users is considered by [21] where green energy is employed in order to migrate VMs to cloudlet and assures a minimal latency between users.

Power consumption for edge IoT is another issue facing this approach of communication. Hence, authors in [22] has proposed various techniques in smart devices in order to prolong the energy life of edge nodes. Moreover, energy optimization is considered in [23] where profiling scheme is proposed in order to monitor the usage trends in mobile applications which would consequently lead to power optimization. On the other hand, mobile devices are equipped with numerous amount of applications to serve different purposes, these apps consume high utilization of battery which affect hereafter the overall power consumption. Therefore, various techniques were proposed in [24] to ensure mobile optimization at the network edge. Additionally, mobile efficiency is discussed in [25] where software defined network is introduced to enhance the overall efficiency by considering two different components such as security and storage. Moreover, power optimization was also reviewed by [26] where a mathematical modelling was proposed to enhance the energy in mobile cloud but in a very large scale network. In addition to previous studies, authors in [27] has proposed a model to place various cloudlets across IoT in order to reduce the average end-to-end delay among end users. However, in line with their study, another work was proposed by [28] where they have introduced different techniques in order to place cloudlets in strategic locations with the aim to minimize the latency but with consideration of balancing the load among cloudlets to avoid overwhelming any of these cloudlets from failures which may lead to a collapse of network.

In contrast to the previous reviewed studies, this paper looks at the edge IoT efficiency and load distribution from network science perspective. Hence, network centrality plays a vital role on load distribution over the network. Therefore, the proposed algorithm defines carefully the nodes centrality at the edge, not only that but also load distribution is enhanced by introducing a reshuffling mechanism which ensures that nodes are not fully occupied by huge connections with other nodes. Thus, this work is in line with previous studies but with new techniques of combining two different science together to enhance edge IoT efficiency. The following section shall give more details of the proposed algorithm.

III. Proposed Method

Mobile edge computing is gaining more attention recently since it's the medium of current communication in most networking environment [29]. Therefore, mobile devices or scattered sensors, which are attached to devices, are forming such type of mobile edge network. This new way of communication has been utilized for solid purpose as discussed in related work section, which is mainly to avoid remote communication with cloud and resides servers on remote areas. However, due to some limitations in the current edge IoT environment, this paper proposes an algorithm, which combines, network science and computer network together to achieve the best optimum efficiency in edge IoT network. As mentioned earlier, mobile devices at the edge of network interact with each other which in returns contribute to the network and help IoT at the edge of not contacting the server for computation and content retrieval. However, since these devices and sensors contribute their power, bandwidth, budgeting, and computing resources. It is very logical to face issues on how to maintain these resources and assure sustainability and stability of the network. Moreover, dealing with high demand on traffic and data would lead to overload of sensors and devices at the edge, this shall cause depletion to these nodes and network failures.

However, different from other previous discussed work, this algorithm is introduced to mainly look at efficiency issue at edge IoT. Therefore, a network science methodology to enhance the load distribution and power efficiency is proposed. Thus, from network science perspective, network centrality metric [30] was the key to measure the efficiency of a dense network at the edge. Thereafter, the communicability [31], which measures the interconnectivity among nodes at the network edge is gauged. In return, after evaluating the communicability values for those nodes located at the edge, the algorithm takes an additional step; to rank these nodes based upon the communicability values in a top-down approach.

Subsequently, the algorithm re-measures the network centrality on those nodes interacting with each other at the edge and keeps reshuffling in order to distribute the load over the these nodes with solid centrality over them. However, the proposed algorithm is also considering the overall degree and closeness on examined nodes. These two metrics play a vital role on the locality of selection of nodes at the edge. Hence, it helps on maintaining nearby nodes instead of interacting with remote devices where traffic travers core network to reach up to destination. This does not only degrade the bandwidth but also affects the overall quality of delivered content. Another critical metric, which could be affected by neglecting the closeness of nodes at the edge, is end-to-end delay due to the random connection among the nodes. Following section shall give more details of implementation of this algorithm where various techniques are combined and introduced together to show the best optimum values of gained efficiency of the examined network.

IV. Evaluation Method

In order to simulate the proposed algorithm, the Python[1] has been adapted as a platform to build and implement a simulated an edge IoT scenario. It is also a high level language with various options of embedded libraries which gives a strong platform for network simulation. However, experimental work has been run over various network sizes randomly although nodes at the edges were carefully selected for execution of the proposed algorithm. Moreover, there are different types of models for topologies, therefore, The Erdős–Rényi model[2] was used to generate the graphs. Hence, at the initial phase, N nodes are generated with full distribution across nodes with budget for every node. At the startup phase, nodes are formed to represent a graph, which shows edge nodes. Afterwards, the algorithm runs a calculation for communicability to gauge the network centrality. This procedure gives a clear image of the conditions of nodes across network and shows the level of interconnection among them.

Thereafter, since the objective of this paper is to focus on edge nodes, the algorithm considers only the values of those nodes. Therefore, the algorithm organizes the edge nodes values as high and low values, hence, the communicability is measured regularly after each run, and values are ordered accordingly. Thus, the gained efficiency is calculated for these nodes and then carefully checks the edge nodes efficiency.

Furthermore, after a thorough analysis of the edge nodes values in terms of communicability and gained efficiency, the algorithm executes a reshuffling scenario, which is known as "rewiring". The aim of this operation is to enhance the load distribution by redirecting links from those edge nodes with high values to those nodes achieving low communicability values. However, It is essential to emphasize that this action takes place iteratively followed by a measurement of gained efficiency to gauge the level of load distribution among edge nodes. However, to ensure that the algorithm does not continue in a loop of reshuffling, an exit condition is defined. This condition is embedded in the algorithm where communicability values are not changing significantly anymore. Moreover, for validation purpose, the algorithm has been run over various sizes of nodes in order to examine the robustness of the algorithm over different conditions.

1) Network centrality: Network centrality is the main function, which gives an insight of the complexity of any type of network. Hence, it has been introduced in the proposed algorithm as of gauging the centrality metric [30]. It reflects as one of the main reference to measure the efficiency of any type of a dense network. Moreover, it captures the level of interconnection between entities, nodes, persons, and so on. In IoT, this plays a vital role where nodes are spread all over. Therefore, it is critical to measure the degree of communication between these nodes and pinpoint carefully the best value of centrality in the network.

2) Communicability: Another factor considered in the proposed algorithm is known as the communicability[3]. This refers to measuring the interconnectivity between nodes all over the network topology. In return, after evaluating the communicability values for all nodes, an additional step is taken by the algorithm, to rank the nodes in the network based upon the communicability, from top to low values.

3) Reshuffling: A new element introduced in the algorithm is "reshuffling". This mechanism acts on re-distributing the links among nodes according to the output of above (A&B). Basically, in some network scenarios, it can be found that some edge nodes are overwhelmed by the amount of outgoing connections whereas other edge nodes are almost at level of depletion cause of having no activity in the network and not contributing the overall distribution. Hence, the main target of this technique is to balance the amount of connections among the edge nodes, taking in consideration that each edge node has at least one connection in the network and staying alive across the network life span. However, a trade-off between reshuffling and network centrality is considered carefully throughout the whole implementation of the algorithm. On the other hand, Quality of service is considered where the algorithm keeps checking packet loss, end-to-end delay metrics to make sure that the minimum level of QoE (Quality of Service) are meeting the minimum acceptable threshold, according to [32].

V. Results and Discussion

This section gives more insights on the output of the simulation of proposed algorithm. Therefore, in order to gauge the performance properly, two scenarios have been introduced. The first scenario represents the proposed algorithm (referred to as "with reshuffling"); whereas the second scenario was introduced to mimic a randomized scenario without any kind of self-adaptation of the WSN (referred to as "without reshuffling"). Moreover, various metrics have been introduced to show the strength of algorithm such as communicability, gained efficiency, closeness, degree, computational time, and number of iterations. These metrics are defined as follows:

[1] "Python." [Online]. Available: https://www.python.org

[2] "Erdos Renyi models.pdf." https://en.wikipedia.org/wiki/Erdős–Rényi_model

[3] "Communicability." [Online]. Available: https://networkx.github.io/documentation/networkx-1.10/reference/algorithms.centrality.html.

Communicability: This metric gives a clear understanding of the load distribution over the WSN in terms of power and energy. The lower values achieved in communicability, the better the load distribution is realized over the network.

Gained Efficiency: This metric is to show network efficiency and how the nodes are contributing to the overall network. It is interlinked with the overall communicability each time. In the proposed algorithm, gained efficiency (GE) is defined as follows:

GE = (Δ/max values) * 100

Where Δ is (MAX - the overall communicability value). The Max value in this case is 1 (depending on the metrics scale).

Degree: This metric shows of how many connections each node at the edge has. It gives an image of the performance of the proposed method in distributing the load at the edge, and helps on alleviating edge node from traffic overload.

Closeness: Since the nodes are spread at the edge, It is essential that nodes connect closely with those nodes close to each other. Therefore, since the proposed algorithm reshuffles the nodes and re-connects them to those nodes with low values of communicability, this procedure helps on distributing the load across edge.

Computational time: This metric is very valuable in measuring the robustness of the proposed algorithm against the randomised approach. Thus, as the network size goes up, this has drastic impact on the performance of showing complicated scenarios. Dealing with a large-scale network size has a significant impact on the performance of executing very complicated scenarios.

Fig. 1 indicates how the network load is distributed across edge IoT network. Thus, both scenarios have been run at the same conditions in terms of network size and connections. However, network graph is randomly generated at the start up, then, communicability is measured for those nodes at the edge only, afterwards, the proposed algorithm starts its process by reshuffling nodes per communicability values. Therefore, it is very obvious from Fig. 1 that proposed method is capable to achieve a steady decrease across various network size which supports the robustness of the proposed algorithm regardless of the network size. Moreover, both scenarios were repeated many times, the results shown here is the average and standard deviation of these runs. This gives meaningful and valid output of the examined scenarios. However, in terms of the performance of the proposed algorithm, It is very clear that communicability values across network size is showing steady decrease and this is an evidence that load and power is fully distributed among nodes at the edge. Therefore, results confirm that such behavior would help on supporting the trend of relying heavily on edge IoT instead of connecting with main servers, which has some limitations as highlighted earlier.

On the other hand, Fig. 2 shows another angle of testing the proposed algorithm, which is the gained efficiency. This metric depends heavily on the performance of the communicability, in other words, it is very linked with each other. Thus, it would not be possible to examine any proposed scenario separately from each other. Therefore, the results, which are achieved here, are consistent with the output highlighted in Fig. 1. Hence, both scenarios have been examined at the same conditions in terms of network size. However, It is obvious that the proposed algorithm is showing steady increase in efficiency, and by looking closely to Fig. 1, It can be claimed that this results is against the achieved communicability by the same algorithm which is an evidence that the lower communicability values are, the more efficiency gained accordingly. Furthermore, results give another line of confirmation that distributing the load at the edge network would be powerful to rely heavily on nodes at the edge, and saving energy and power since they contribute by their own computing resources. However, looking at the randomized scenario which is called “without reshuffling”, It is obvious that gained efficiency is almost obsolete due to the randomness in connections, not only that but also, load is not distributed among nodes, and this is backed up by the results shown in Fig. 1 where the communicability values are steady line across various network size.

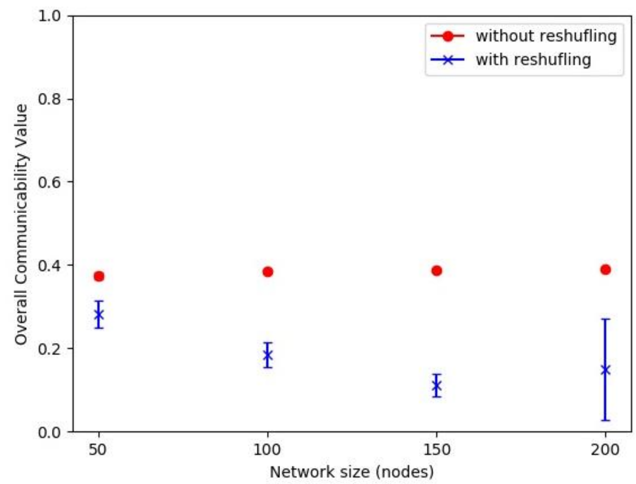


Fig. 1. Overall Communicability.

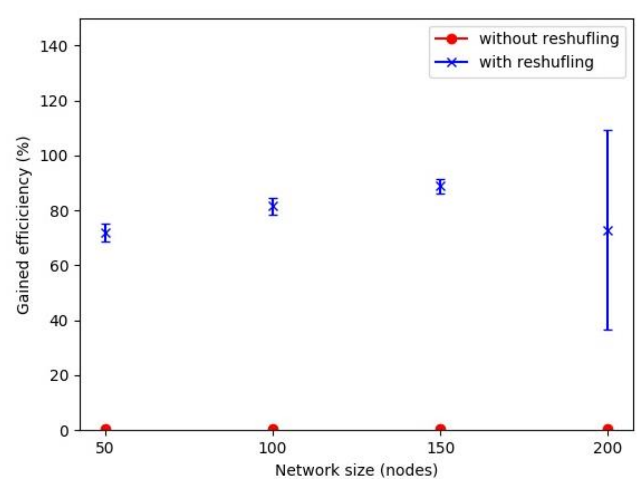


Fig. 2. Gained Efficiency.

Another factor that plays a vital role on edge IoT network is the degree value. This metric has been selected to show the amount of connections that each node at the edge is handling. Hence, the lower connections mean that the higher load distribution is achieved across the edge. Therefore, Fig. 3 shows the degree level at the proposed algorithm "with reshuffling" against the scenario "without reshuffling". Subsequently, the results show that the embedded techniques have shown a steady decrease in degree level, and this is consistent with the previous highlighted results in Fig. 1 and 2. On other hand, the benchmarking scenario shows steady line where the number of connections at each run over various network size is almost the same. Thus, there is no consideration of load distribution as well as energy efficiency, which could affect the nodes at the edge and lead to battery depletion, which may incur network failure.

Furthermore, Fig. 4 gives another insight of the value of the proposed algorithm, which measures the closeness among connected nodes at the edge level. Hence, Fig. 4 shows a clear evidence that combining network science and computer network not only enhance the load distribution and efficiency but helps on the closeness among nodes. Hereafter, Fig. 4 indicated that the overall closeness is decreased steady against the network size, this is certainly an evidence that the proposed algorithm is performing efficiently across various conditions.

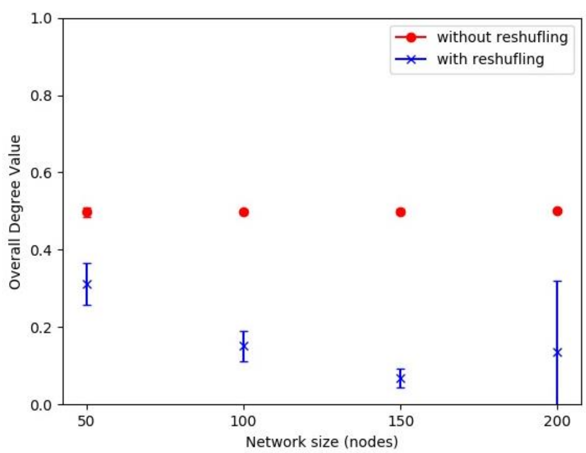


Fig. 3. Overall Degree.

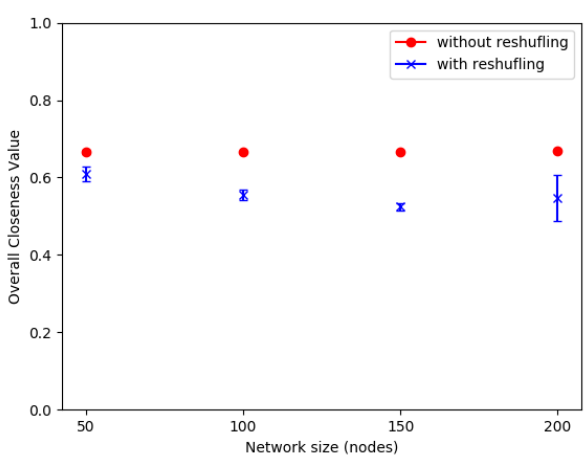


Fig. 4. Overall Closeness.

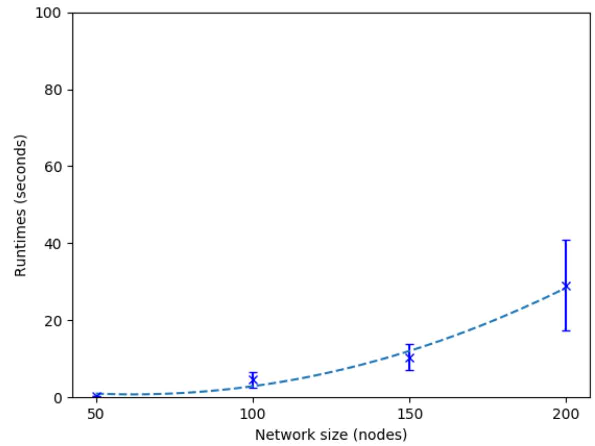


Fig. 5. Computational Time.

Computational time is another critical factor of the proposed algorithm, since it shows the robustness of the proposed approach. Therefore, dealing with a large-scale network size has a significant impact on the performance of executing very complicated scenarios. This metric was very valuable in measuring the computational efficiency of the proposed algorithm, in terms of running time and execution.

Fig. 5 indicates that the running time increases linearly against the network size. A careful examination of the amount of elapsed time necessary to run, execute, and achieve the best results over many runs, demonstrates that the algorithm performs very well despite the network size and the complexity of connections among the nodes.

VI. Conclusion

Edge IoT is given more attention recently due to the advantages shown against traditional IoT paradigm, which relies heavily on connecting to reside servers at the main core network. Therefore, this paper presents an edge IoT scenario with combination of network science and computer network in order to show how such combination would lead to efficient utilization of edge IoT network. Results have clearly shown that edge IoT is very promising approach in IoT, with the consideration of load distribution and power efficiency. Hence, the proposed algorithm has shown robustness in terms of various metrics such as network centrality, communicability, degree, and closeness. Moreover, the proposed algorithm was compared to mimic a traditional scenario, which has no techniques to enhance the efficiency and, load distribution. Future work will focus more on QoS and QoE metrics by injecting a video streaming across network and examine the proposed algorithm over various condition of network.

References

[1] J. Ren, Y. Zhang, R. Deng, N. Zhang, D. Zhang, and X. S. Shen, "Joint channel access and sampling rate control in energy harvesting cognitive radio sensor networks," IEEE Trans. Emerg. Top. Comput., vol. 7, no. 1, pp. 149–161, 2019.

[2] T. Saarikko, U. H. Westergren, and T. Blomquist, "The Internet of Things: Are you ready for what's coming?," Bus. Horiz., vol. 60, no. 5, pp. 667–676, 2017.

[3] M. Gayathri and K. Srinivas, "A Survey on mobile cloud computing architecture, applications and challenges," Int. J. Sci. Res. Eng. Technol., vol. 3, no. 6, pp. 1013–1021, 2014.

[4] P. Bellavista, J. Berrocal, A. Corradi, S. K. Das, L. Foschini, and A. Zanni, “A survey on fog computing for the Internet of Things,” Pervasive Mob. Comput., vol. 52, pp. 71–99, 2019.

[5] A. Pal and K. Kant, “NFMI: Connectivity for Short-Range IoT Applications,” Computer (Long. Beach. Calif)., vol. 52, no. 2, pp. 63–67, 2019.

[6] Q. Wang, D. Lin, P. Yang, and Z. Zhang, “An Energy-Efficient Compressive Sensing-Based Clustering Routing Protocol for WSNs,” IEEE Sens. J., vol. 19, no. 10, pp. 3950–3960, 2019.

[7] R. Yarinezhad, “Reducing delay and prolonging the lifetime of wireless sensor network using efficient routing protocol based on mobile sink and virtual infrastructure,” Ad Hoc Networks, vol. 84, pp. 42–55, 2019.

[8] M. Chincoli and A. Liotta, “Self-learning power control in wireless sensor networks,” Sensors (Switzerland), vol. 18, no. 2, pp. 1–29, 2018.

[9] A. A. Banka and R. N. Mir, “Current Big Data Issues and Their Solutions via Deep Learning: An Overview.,” Iraqi J. Electr. Electron. Eng., vol. 14, no. 2, pp. 127–138, 2018.

[10] M. Amadeo, G. Ruggeri, C. Campolo, and A. Molinaro, “IoT services allocation at the edge via named data networking: From optimal bounds to practical design,” IEEE Trans. Netw. Serv. Manag., vol. 16, no. 2, pp. 661–674, 2019.

[11] Y. Ai, M. Peng, and K. Zhang, “Edge computing technologies for Internet of Things: a primer,” Digit. Commun. Networks, vol. 4, no. 2, pp. 77–86, 2018.

[12] M. Satyanarayanan, “The emergence of edge computing,” Computer (Long. Beach. Calif)., vol. 50, no. 1, pp. 30–39, 2017.

[13] A. M. Khan and F. Freitag, “On Participatory Service Provision at the Network Edge with Community Home Gateways,” Procedia Comput. Sci., vol. 109, pp. 311–318, 2017.

[14] S. Kim, “Nested game-based computation offloading scheme for Mobile Cloud IoT systems,” Eurasip J. Wirel. Commun. Netw., vol. 2015, no. 1, 2015.

[15] B. Li, Z. Fei, J. Shen, X. Jiang, and X. Zhong, “Dynamic Offloading for Energy Harvesting Mobile Edge Computing: Architecture, Case Studies, and Future Directions,” IEEE Access, vol. 7, pp. 79877–79886, 2019.

[16] Q. Fan and N. Ansari, “Cost Aware cloudlet Placement for big data processing at the edge,” IEEE Int. Conf. Commun., pp. 1–6, 2017.

[17] F. Slim, F. Guillemin, and Y. Hadjadj-Aoul, “On virtual network functions’ placement in future distributed edge cloud,” Proc. 2017 IEEE 6th Int. Conf. Cloud Networking, CloudNet 2017, pp. 1–4, 2017.

[18] L. A. Tawalbeh, W. Bakheder, and H. Song, “A Mobile Cloud Computing Model Using the Cloudlet Scheme for Big Data Applications,” Proc. - 2016 IEEE 1st Int. Conf. Connect. Heal. Appl. Syst. Eng. Technol. CHASE 2016, pp. 73–77, 2016.

[19] M. Satyanarayanan et al., “Edge analytics in the internet of things,” IEEE Pervasive Comput., vol. 14, no. 2, pp. 24–31, 2015.

[20] X. Sun and N. Ansari, “EdgeIoT: Mobile Edge Computing for the Internet of Things,” IEEE Commun. Mag., vol. 54, no. 12, pp. 22–29, 2016.

[21] X. Sun, N. Ansari, and Q. Fan, “Green energy aware avatar migration strategy in green cloudlet networks,” Proc. - IEEE 7th Int. Conf. Cloud Comput. Technol. Sci. CloudCom 2015, pp. 139–146, 2016.

[22] K. S. Bahwaireth, L. Tawalbeh, A. Basalamah, Y. Jararweh, and M. Tawalbeh, “Efficient techniques for energy optimization in Mobile Cloud Computing,” Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA, vol. 2016-July, pp. 1–8, 2016.

[23] E. Benkhelifa, T. Welsh, L. Tawalbeh, Y. Jararweh, and A. Basalamah, “User profiling for energy optimisation in mobile cloud computing,” Procedia Comput. Sci., vol. 52, no. 1, pp. 1159–1165, 2015.

[24] L. A. Tawalbeh, A. Basalamah, R. Mehmood, and H. Tawalbeh, “Greener and Smarter Phones for Future Cities: Characterizing the Impact of GPS Signal Strength on Power Consumption,” IEEE Access, vol. 4, pp. 858–868, 2016.

[25] L. Tawalbeh, Y. Haddad, O. Khamis, F. Aldosari, and E. Benkhelifa, “Efficient software-based mobile cloud computing framework,” Proc. - 2015 IEEE Int. Conf. Cloud Eng. IC2E 2015, pp. 317–322, 2015.

[26] M. Al-Ayyoub, Y. Jararweh, L. Tawalbeh, E. Benkhelifa, and A. Basalamah, “Power Optimization of Large Scale Mobile Cloud Computing Systems,” Proc. - 2015 Int. Conf. Futur. Internet Things Cloud, FiCloud 2015 2015 Int. Conf. Open Big Data, OBD 2015, pp. 670–674, 2015.

[27] Z. Xu, W. Liang, W. Xu, M. Jia, and S. Guo, “Efficient Algorithms for Capacitated Cloudlet Placements,” IEEE Trans. Parallel Distrib. Syst., vol. 27, no. 10, pp. 2866–2880, 2016.

[28] M. Jia, J. Cao, and W. Liang, “Optimal Cloudlet Placement and User to Cloudlet Allocation in Wireless Metropolitan Area Networks,” IEEE Trans. Cloud Comput., vol. 5, no. 4, pp. 725–737, 2015.

[29] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, “A Survey on Mobile Edge Computing: The Communication Perspective,” IEEE Commun. Surv. Tutorials, vol. 19, no. 4, pp. 2322–2358, 2017.

[30] S. Wang, Y. Du, and Y. Deng, “A new measure of identifying influential nodes: Efficiency centrality,” Commun. Nonlinear Sci. Numer. Simul., vol. 47, pp. 151–163, 2017.

[31] E. Estrada, “The communicability distance in graphs,” Linear Algebra Appl., vol. 436, no. 11, pp. 4317–4328, 2012.

[32] F. Agboma, M. Smy, and A. Liotta, “QoE analysis of a peer-to-peer television system,” MCCSIS’08 - IADIS Multi Conf. Comput. Sci. Inf. Syst. Proc. Wirel. Appl. Comput. 2008 Telecommun. Networks Syst. 2008, pp. 114–119, 2008.

Awareness of Ethical Issues when using an e-Learning System

Talib Ahmad Almseidein[1], Omar Musa Klaif.Mahasneh[2]
Department of basic and Applied Science, Shoubak University College
AL-Balqa Applied University, AL-Salt, Jordan

***Abstract*—Transformation to the digital system has made life easier, and the acceptance of the e-learning system in the academic life of students is a fact. Therefore, many educational organizations use the e-learning environment for teaching-learning activities. The present study has been conducted to evaluate the awareness of the undergraduate students to ethics and to determine if there is a difference according to gender and academic level variables when using an e-learning system at Al-Balqa Applied University in Jordan. A self-questionnaire has been designed to measure the participant's awareness of ethics. It consists of 20 items classified in three ethical categories; Intellectual property rights, vandalism and Privacy. The results show that the awareness of students is low in all three categories regarding their commitment to the ethical issues when using an e-learning system. Result also show that there are no significant differences between undergraduate students' gender and academic level related to the awareness of ethical issues. Therefore, undergraduate students should be fully knowledgeable about ethical issues to avoid unethical behavior while using of the e-learning system.**

Keywords*—*Code of ethics; ethical issues; e-learning

I. Introduction

Technological development at this current stage requires university students to use the computer and the Internet to gain knowledge. This has made it easier for university students to access many sources of information that can be used in education. Through advanced technology, university students can submit assignments, send files, and access necessary academic information and other services. Despite the benefits of using technology for students, it can increase the chances of falling into unethical behaviors [1].

The Education community has a belief that the traditional teaching-learning models do not meet the new challenges created by coming out technologies [2]. Therefore, many educational organizations turned towards the adoption of Information Communication Technology (ICT) to enhance their functions and goals. Nowadays, the quick growth of technology has changed the environment where life is affected by technology. The teaching-learning process relies on the employment of ICTs to achieve the educational goals by using e-learning [3]. Author of [4] stated that learning by facilitating and supporting through using ICT is an E-learning. Activities as managing data, storing information and sharing materials allow students and teachers to communicate and handle the assignments electronically and monitor the course through an e-learning system [5]. Besides, the distribution of study substances, evaluation of the capabilities of students and the enhancement of students through interaction between student and teacher is achieved through the use of ICTs in e-learning institutions. Recently, academic institutions are using e-learning as an educational platform called Learning Management System (LMS) to its benefits as asynchronous and synchronous learning, improved communication and more collaboration [6]. Academic institutions use two modes of e-learning. The first one is the Online mode where all activities are done through technology without using physical interaction between learner and instructor. The second one is the integration of face-to-face and online environments, known as a blended model [7]. In spite of e-learning advantages, it has aggravated the problems of cheating, plagiarism, and violation of privacy [8]. Furthermore, computer abuses, computer crimes, theft of data, equipment malfunctions, destruction from viruses, errors in handling and usage of data are ethical issues, which are considered to be threats to e-learning [9].

Professional international computer groups such as the (ACM, IEEE, and others) have developed written codes of ethics and make them obligatory for the members to follow. This interest is proof of the importance of understanding ethics [10].

This study aims to evaluate the student’s awareness of ethics when using an e-learning system for the teaching-learning process. It explores if there is a difference in the awareness of the ethical issue when using the e-learning system according to gender and academic level variables. The paper is divided into six sections excluding the introduction section. Section II discusses the literature review. The research methodology is revealed in Section III. Section IV shows the result. Section V tackles the discussion, while the conclusion and recommendations are discussed in Section VI.

II. Literature Review

Chances for unethical behavior through students usually take place in e-Learning rather than traditional learning, which led the learners in e-learning to behave unethically [11, 12]. Ethical behavior is socially and morally acceptable, which is correspondent with right doings of the society [13]. In other words, ethics is a guide to making decisions about what is right and what is wrong.

Author in [14] has done a quantitative study to investigate the ethical issues that are related to the use of ICT by students at a higher education institution in the Eastern Capital province of South Africa. The number of undergraduate students was 312 and they sampled with the use of a questionnaire. The survey highlighted plagiarism, Software piracy, and cheating impact issues. The author concluded that ethical issues have to be covered in the curriculum when the author found that students were not aware of ethical issues as plagiarism and software piracy.

Author in [15] has proposed a model to integrate ethics in e-learning through virtual academic Counsellor to improve the ethical values of students and suggested recommendations to be followed which will help to produce ethically qualified graduates.

Authors of [7] indicated that Through the respondents' opinion in online courses the majority of students using the e-learning system conducted unethical behavior as cheating. Authors of [8] have examined cheating behaviors such as plagiarism among 47 undergraduate education final years students. They concluded that using e-learning is exacerbating the problem of cheating.

Author in [3] has done a study by using two questionnaires to study the code of ethics of teaching and learning for an e-learning system. The results indicated that teachers and students have to be aware of ethics and its future ramifications when using the e-learning system and activities offered by its demand to be advanced. Moreover, many researchers mentioned academic fraud is a very common ethical issue in e-Learning, which is including assistance on examinations, plagiarism and the absence of an obligation to copyright [16-18].

According to The National Survey of Student Engagement (NSSE), 59% of the United States students used e-learning programs have reported that they did some sort of academic fraud in their study, 27% did it very often while 32% often. [19].

III. Methodology

In this paper, a self-questionnaire is used as a research instrument. The questionnaire has one part which evaluates the student's awareness of ethics when using an e-learning system. It contains 20 items, which have been obtained from the code of ethics. The code of ethics is the standard of behavior that guides decisions and actions, based on duties that have been derived from core values [20]. In this study, the authors use the principles reported by (ACM and IEEE) to evaluate the student's awareness of ethics. The authors categorize the 20 items into three categories; Intellectual property rights, Vandalism and Privacy. A five-point Likert Scale has been used ranging from 5 (strongly agree), 4 (agree), 3 (neutral), 2 (disagree) and 1 (strongly disagree), based on the main items.

The collected data are processed and statistically analyzed by using (SPSS) Version 22. Range of means and their standard deviations values are used to evaluate the student's awareness of ethics when using an e-learning system. Larger values of the mean indicate higher levels of ethical awareness and lower values implied lower levels of ethical awareness. The strength of the ethical awareness is categorized into three levels based on the values of means. The authors adopt the following standard to indicate the degree of students' awareness of electronic ethics while using e-learning. There is low, moderate and high when the mean value is ≤ 2.32, $3.66 - 2.33$ and ≥ 3.67, respectively. Range = (5-1) /3=1.33.

The reliability factor for the questionnaire is 0.85 which indicates a good level of reliability [21]. The questionnaire study has been applied to a random sample of 108 undergraduate students at Al-Balqa applied university (BAU). The BAU is one of the largest public universities in Jordan and has integrated e-learning in their degree program.

IV. Result

The objective of this study is to evaluate the student's awareness of ethics and if there is a difference according to gender and academic level variables when using an e-learning system at the BAU for the teaching-learning process. The random study sample consists of 108 undergraduate students, their specializations and degrees are different. The awareness of students is grouped into three categories: Intellectual Property Rights, Vandalism and Privacy. The results of Means, standard deviations and awareness of students for each element measured in this study as shown in Tables I-III.

To determine if there is a difference awareness of ethical issue when using the e-learning system according to gender and academic level variables, ANCOVA variance analysis has been used and results shown in Table IV.

The Intellectual Property Rights consists of eight items. The items of this area are about copying and documentation. The values in Table I show that student’s awareness is low for each item in this category.

TABLE. I. Means ±SD of Intellectual Property Rights Category

Number	Item	Mean ±SD	Awareness
1	I accept to copy my colleague's duties and research without his approval and send it through the e-learning system.	1.18±.383	Low
2	Doing the assignments without documenting references and send them through the e-learning system.	1.69±.769	Low
3	Preparing research and reports without documenting references and send them through the e-learning system	1.81±.767	Low
4	Use software applications without prior permission from the original product to do the assignments and upload them through the e-learning system	1.69±.606	Low
5	Use ready-made presentations from the Internet without documenting the reference and attribute it to me and send it through the e-learning system	1.67±.641	Low
6	Photocopy documents from references and send them through the e-learning system without the consent of the publisher	1.64±.703	Low
7	Ask others to get their homework solutions and attribute it to me then send it through the e-learning system	1.65±.701	Low
8	Use images and diagrams on the Internet without documenting the reference	1.34±.751	Low
Total mean of this category = 1.58			

Table II shows the results for the Vandalism category which consists of seven items, related to acts to harm others such as sending malicious files or destroying others' data. The values show that students' awareness is low for each item in this category.

The last category results are showed in Table III, as well as the values of each item in this category is low to determine the student's awareness.

Analysis of (ANCOVA) test has been applied to determine if is there a difference in the awareness of ethical issues when using the e-learning system according to gender and academic level variables. A significant level of $\alpha < 0.05$ is adopted for the study. As shown in Table IV, there are no significant differences between undergraduate students' gender and academic level about awareness of ethical issues.

TABLE. II. MEANS ±SD OF VANDALISM CATEGORY

Number	Item	Mean ±SD	Awareness
1	Send files filled with viruses through the e-learning system.	1.56±.535	Low
2	Send files through the e-learning system without making sure they are safe.	1.14±.398	Low
3	Access the e-learning system from other accounts without their permission.	1.98±.192	Low
4	Open files via the e-learning system without making sure they are safe	1.76±.609	Low
5	Send immoral files or links through the e-learning system.	1.99±.837	Low
6	Send applications through the e-learning system to destroy other devices	1.66±.686	Low
7	I try to encrypt the data of others for barter	1.64±.633	Low
Total mean of this category = 1.67			

TABLE. III. MEANS ±SD OF PRIVACY CATEGORY

Number	Item	Mean ±SD	Awareness
1	I give my colleagues the login information for my account in the e-learning system	1.71±.938	Low
2	Enter my colleagues' e-learning accounts to solve their jobs and send them through the e-learning system	1.47±.502	Low
3	Publish the files related to the educational material in the e-learning system via social networks	1.44±.584	Low
4	I preserve the files of others that I found when using the chat in the e-learning system	1.76±.594	Low
5	I use the assignment solution of my old colleagues and send them through the e-learning system without their permission	1.76±.668	Low
Total mean of this category = 1.62			

TABLE. IV. COMPARISON OF ETHICAL ISSUES AWARENESS WHEN USING E-LEARNING SYSTEM BY GENDER AND ACADEMIC LEVEL

Dependent Variable: Ethics					
Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	45.401a	7	6.486	.763	.619
Intercept	46442.272	1	46442.272	5.466E3	.000
Gender	8.111	1	8.111	.955	.331
Level	16.644	3	5.548	.653	.583
gender * level	2.837	3	.946	.111	.953
Error	849.590	100	8.496		
Total	115035.000	108			
Corrected Total	894.991	107			
a. R Squared = .051 (Adjusted R Squared = -.016)					

V. DISCUSSION

Results show that the level of awareness of the issues associated with ethics when using an e-learning system by students is low. The results show that the Category of "Intellectual Property Rights" obtains the mean value (1.58); the second category "Vandalism" with a mean value of (1.67). The last mean value is (1.62). The results of this study, which has been conducted on a random sample of the students at BAU, indicate that the awareness of students of their commitment to the ethics when using an e-learning system comes low in all categories. This indicates that the students have unethical awareness about the discussed issues these categories. In light of the study results, it has been found that there are ethical issues among students when using e-learning, and this is confirmed in the study of [22] who highlighted the existence of ethical issues when using e-learning.

The authors believe that the reasons for this comes result of several factors, such as the difference in students' backgrounds and culture to determine what is ethical or immoral in the e-learning environment, lack of include ethics within the curricula, the digital division and the lack of a general policy from academic institutions clarifying the ethics of dealing with the e-learning environment.

The result also indicates that there are no significant differences between the undergraduate students' gender and academic level regarding their awareness of ethical issues. Where the level of significances (0.331,0.583) is higher than the level of significance (0.05).

VI. CONCLUSION AND RECOMMENDATIONS

The results of this study reveal that students at BAU have ethical issues when using e-learning system. To use e-learning system by the students within their academic period, the students should be fully knowledgeable about ethical issues to avoid unethical behavior. Authors of [23] discussed the inclusion of professional ethics topics in courses and developing e-learning on professional ethics so that students get an opportunity to understand and practice ethical judgments in the various aspects of ICT work. On the other hand, authors in [24] reported there is a need for academic institutions to espouse guidelines and ethical policies for the e-Learning environment for preserving and supporting academic integrity.

The following are some recommendations for the institutions delivering e-Learning systems based on the results of the study:

- The inclusion of the subject of the ethical values as a seminar course within the student academic plan.
- The educational institution must build proper rules and policies for students involved in e-learning.
- The educational institution should adopt multimedia tools and monitoring tools in e-learning environment.
- Workshops or training must be held about ethical issues for students.
- Putting the necessary instructions on the optimal use of the e-learning system when student access his account to avoid any unethical behavior.
- Academic institutions should adopt a Code of ethics for their students when using the e-learning system seriously.

Finally, the study recommends the need to do similar studies on other universities in Jordan, intending to increase awareness and knowledge of ethical issues related to e-learning system.

References

[1] Jamil, M., and Shah, J. H. "Perception of undergraduates about computer and Internet ethics in Pakistan," Nigerian Journal of Technology (NIJOTECH), 33(4), 512-522, 2014.

[2] Nagi, K. "Solving Ethical Issues in eLearning," Special Issue of the International Journal of the Computer, the Internet and Management, 14(SP1), 2006.

[3] AL-Shehri, M. "Code of Ethics of Teaching-Learning for an e-Learning System," International Journal of Computer Applications, 166(5), 16-20, 2017.

[4] Masrom, M. "Technology Acceptance Model and E-learning," 12th International Conference on Education, Sultan Hassanal Bolkiah Institute of Education Universiti Brunei Darussalam,2017.

[5] Lonn, S., and Teasley, S. "Saving time or innovating practice: Investigating perceptions and uses of Learning Management Systems," Computer & Education, 53(3), 686-694, 2009.

[6] Nawaz, A., and Kundi, G. M. "Demographic implications for the user-perceptions of e-learning in higher education institutions of NW. FP," PAKISTAN, Electron. J. Inf. Syst. Dev. Ctries., 41(5), 1-17, 2010.

[7] Muhammad, A., Ghalib, M.F.MD., Ahmad, F., Naveed, Q.N., & Shah, A. "A Study to Investigate State of Ethical Development in E-Learning," (IJACSA) International Journal of Advanced Computer Science and Application, 7(4), 284-289, 2016.

[8] Trushell, J., Byrne, K., and Simpson, R. "Cheating behaviours, the Internet and Education undergraduate students," Journal of Computer Assisted Learning, 28(2), 136–145, 2012

[9] Eftekhari, S. R. "e-Learning and e-Ethics," The Eighth International Conference on eLearning for Knowledge-Based Society, 23-24 February 2012, Thailand.

[10] Wood, G. "A cross cultural comparison of the contents of codes of ethics: USA, Canada and Australia," J. Bus. Ethics, . 25(4), 287–298, 2000.

[11] AbdulHafeez, M., Farooq, A., and Asadullah, R. "Resolving Ethical Dilemma in Technology Enhanced Education through smart mobile devices," Int. Arab J. e-Technology, 4(1), 25–31, 2015.

[12] Muhammad, A. H., Wahsheh, H. A., Shah, A., and Ahmad, F. "Ethical perspective of learning management system a model to support moral character of online learner," in Information and Communication Technology for The Muslim World (ICT4M), The 5th International Conference on, 2014, pp. 1–6, 2014.

[13] Bowden, P., and Smythe, V. "Theories on teaching & training in ethics," EJBO Electronic Journal of Business Ethics and Organization Studies, 13(2), 19-26, 2008.

[14] Cilliers, L. "Evaluation of information ethical issues among undergraduate students: An exploratory study," South African Journal of Information Management 19(1), a767, 2017.

[15] Muhammad, A. "Integration of ethics in e-Learning through Virtual Academic Counsellor," 2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) , Sakhier, Bahrain, 2018, pp. 1-6.

[16] Fess, R. A. "Cheating and plagiarism. Ethics and Higher Education." May, W. W. editor. New York: Macmillan Publishing Company and American Council on Education, 1990.

[17] Farisi, M. I. "Academic dishonesty in distance higher education: Challenges and Models for Moral Education in the Digital Era," Turkish Online Journal of Distance Education, 14(4), 176-195, 2013.

[18] Melissa, R., Olt, M.A., and M.S. "Ethics and Distance Education: Strategies for Minimizing Academic Dishonesty in Online Assessment," Online Journal of Distance Learning Administration, 5(3), 2002.

[19] Kuh, G. D. "The national survey of student engagement: Conceptual and empirical foundations," In R. Gonyea & G. Kuh (Eds.), Using student engagement data in institutional research (New Directions for Institutional Research Series, no. 141, pp.5–20). San Francisco: Jossey-Bass, 2009.

[20] North, M. M., George, R., and North, S. M. " Computer Security and Ethics Awareness in University Environments: A Challenge for Management of Information Systems." ACM-SE 44 Proceedings of the 44 annual Southeast regional conference Pages 434-439, 2006.

[21] Bryman, A. "Social research methods, Oxford University Press," Oxford, 2012.

[22] Muhammad, A., Ahmad, F., and Shah, A. "Resolving Ethical Dilemma in Technology Enhanced Education through smart mobile devices," International Arab Journal of e-Technology, 4(1), 25–31, 2015.

[23] Nagarajan, K., Bevinakoppa, S. "Towards Understanding the Role of Institutional Pressures in Raising Awareness about Student Academic Integrity Issues in Australia," International Journal of Education and Learning Systems, 3, 93-104, 2018.

[24] A. Muhammad, A. Shaikh, Q. N. Naveed and M. R. N. Qureshi, "Factors Affecting Academic Integrity in e-Learning of Saudi Arabian Universities. An investigation using Delphi and AHP," in IEEE Access. doi: 10.1109/ACCESS.2020.2967499.

A Blockchain based Mobile Money Interoperability Scheme

Fickson Mvula[1], Simon Tembo[3]
Department of Electronics and Electrical Engineering
School of Engineering, UNZA
Lusaka, Zambia

Jackson Phiri[2]
Department of Computer Science
School of Natural Sciences, UNZA
Lusaka, Zambia

***Abstract*—Developing Countries in Africa in general and Zambia in particular, have seen a rapid rise in use of mobile payment platforms. This has not only revolutionized access to finance for the poor but also allowed them access to other financial products such as savings or insurance. With a growing number of mobile money providers in Zambia, there is need for a solution that would enable integration of the mobile money provider's systems using a central clearinghouse for purposes of clearing and settlement to achieve mobile money interoperability. In this study, we first reviewed the technical landscape and features of mobile payment systems in Zambia and then assessed the feasibility of using blockchain technology in proposing a settlement and clearing system that would facilitate mobile money interoperability. A prototype system was then designed in which amounts being interchanged between providers are managed as assets on a permissioned blockchain. The system runs a distributed shared ledger, which provides non-repudiation, data privacy and data origin authentication, by leveraging the consistency features of blockchain technology.**

***Keywords*—Blockchain; mobile money interoperability; clearing and settlement; blockchain security**

I. Introduction

There is a growing number of mobile money wallet services providers in Zambia which has led to the creation different autonomous financial ecosystems with little to no interoperability between them. We define interoperability as an ability of one mobile money subscriber on one network, to transfer value to another on a different network [1]. Attempts have been made to close this gap through provision of bilateral arrangements between mobile money providers which has proved problematic as there are delays in settlement due to ledger trust issues.

Currently, there is currently no live implemented system that allows interoperability between the different mobile financial services wallet providers in Zambia. The proposed Zambia National Switch project [4] being undertaken by the Zambia Electronic Clearing House Limited (ZECHL) will among others enable participants in the mobile financial ecosystem to interchange money by providing a clearing and settlement platform. The system implementation will be phased and the first phase expected to cater for interoperability of commercial banks and expected to be launched at the end of 2019. The second phase will cater for integration of other financial services such as mobile money and telegraphic money transfers [5].

The National Financial Switch system however, being a traditional database based central system will have a number of shortfalls in as far of effective provision of the desired features identified for clearing and settlement of account to account (A2A) interoperability transactions. Firstly, there will be integration complexity as every participant will be required to connect to a central node. This central node of processing will hinder efficiencies in end-to-end processing speed and thus availability of funds may be hampered. Further, there will be no network resilience offered by distributed data management system such as one provided by a distributed ledger system. And furthermore, there may be operational and financial risks as a result of a single central node rather than a distributed one.

Integration of wallet provider's systems through a central clearing house for purposes of clearing and settlements [2] is necessary to achieve interoperability. Blockchain technology presents a perfect opportunity as a potential technology to disrupt payment, clearing and settlement because of its ability to introduce a set of synchronized ledgers managed by one or more entities rather than individual non communicating ledgers [3]. This would lead to a reduction in the reliance on traditional central ledger managed by a trusted entity for holding and transferring funds.

In this paper, we present a study that proposes the design of a secure and trusted blockchain based clearing and settlement architecture that will allow seamless interoperability for mobile financial services in Zambia. The paper is divided into five sections. Section 2 gives a background to Blockchain technology and shows how it could be used to support the use case in the study while Section 3 gives the literature review and describes similar approaches that others have used to solve the interoperability problem. Section 4 presents the research methodology while Section 5 covers the results of the research.

II. Clearing and Settlement on a Blockchain

A. Blockchain Defined

A blockchain can be defined as an immutable ledger for recording transactions, maintained within a distributed network of mutually untrusting peers. Every peer maintains a copy of the ledger. The peers execute a consensus protocol to validate transactions, group them into blocks, and build a hash chain over the blocks. This process forms the ledger by ordering the transactions, as is necessary for consistency. Blockchains have emerged with Bitcoin and are widely regarded as a promising technology to run trusted exchanges in the digital world [6].

The two main categories of blockchains are public and private blockchains. In a public or permission-less blockchain, anyone can participate without a specific identity. Public blockchains typically involve a native cryptocurrency and often use consensus based on proof of work (PoW) and economic incentives. Permissioned blockchains, on the other hand, run a blockchain among a set of known, identified participants. A permissioned blockchain provides a way to secure the interactions among a group of entities that have a common goal but which do not fully trust each other, such as businesses that exchange funds, goods, or information.

By relying on the identities of the peers, a permissioned blockchain can use traditional Byzantine-fault tolerant (BFT) consensus instead of incentives based consensus mechanisms.

B. Blockchain in Funds Clearing and Settlement

The proposed solution is directed primarily at arrangements that involve restricted ledgers (access to which is for approved users only) or permissioned blockchains, reflecting the main types of arrangement currently being developed in the financial sector, such as one required for a mobile money account to account interoperability among a number of disparate network providers.

Clearing and settlement of a financial transaction, regardless of the asset type, requires a network of participants, an asset or set of assets that are transferred among those participants, and a transfer process that defines the procedures and obligations associated with the transaction. Typically, the set of direct participants are financial institutions such as banks or brokers and indeed mobile wallet providers in the case of mobile financial services. Indirect participants include end users such as subscribers in this case. An asset can be any financial instrument, such as a monetary instrument, security, commodity, or a derivative.

In a mobile financial services ecosystem, the asset type of interest is virtual money (or e-money) being transferred from one wallet to another across the network of participants. Communications among the participants in a network involve sending electronic messages, acknowledgements, statements, and other information between computer systems typically maintained by a network operator and its participants.

It is worth noting at this stage that the current implementation of such networks is such that each participant maintains and is responsible for their own financial ledger which acts as their single source of truth on the status of their data. To achieve interoperability, a common central authority may be necessary which would be entrusted by their participants with updating and preserving the integrity of a central ledger and, in some cases, managing certain risks on behalf of participants.

The case for Distributed ledger technology (DLT) as a potential technology to disrupt payment, clearing and settlement implementations is because of the technology's ability to introduce a set of synchronized ledgers managed by one or more entities rather than individual non communicating ledgers. This would lead to a reduction in the reliance on traditional central ledger managed by a trusted entity for holding and transferring funds and other financial assets.

DLT may radically change how assets are maintained and stored, obligations are discharged, contracts are enforced, and risks are managed. Proponents of the technology highlight its ability to transform financial services and markets by [7]:

- Reducing complexity.
- Improving end-to-end processing speed and thus availability of assets and funds.
- Decreasing the need for reconciliation across multiple record-keeping infrastructures.
- Increasing transparency and immutability in transaction record keeping.
- Improving network resilience through distributed data management.
- Reducing operational and financial risks.

III. Literature Review

In their basic sense Mobile payments platforms allow their users to pay and transfer funds in mobile money, but also offer access to other financial products, such as savings and bill payments. A study in [8], reviewed the economic features of mobile payment systems in developing countries, and studied the cooperation models that can emerge between the different firms potentially involved in a mobile payment transaction. Focus was drawn on the main competition concerns that public authorities should be concerned about, and which regulatory tools could be considered as a remedy. Key among some of the key challenges in mobile money schemes was the issue of interoperability. Different concepts of interoperability are relevant and need to be distinguished according to their implications for regulation and business models differ.

Different approaches have been undertaken by different countries in an attempt to implement interoperability for their mobile money financial systems. This section reviews a number of such proposed architectures for mobile payments that support interoperability. These have been drawn from well-developed mobile money markets and they include India, Kenya, Rwanda, and Tanzania. Next, a number of blockchain based use cases were reviewed and presented to support the case for use of blockchain in a system model proposed.

A. Interoperability Schemes in other Similar Markets

In 2008, the Reserve Bank of India (RBI) provided an interoperability platform called UPI [9]. This is however, a central integrating node which suffers integration complexities.

Alternative architecture approaches proposed [10] with hierarchical lookup. Kumar et al. also proposed architectural choices [11]. However, their model is specific to highly regulated financial environment in India, where every transaction is processed by a bank.

Other options in the Indian landscape include, the Mobile Payment Foundation of India [11] which is also developing a model for interoperability. Further, Kumar et al. have proposed architectural choices for interoperability [11]. However, their model is specific to highly regulated financial environment in India, where every transaction is processed by a bank.

Interoperability is not mandated under the Kenyan National Payments System (NPS) regulations but instead payment service use bilateral arrangements [12], [13], [14] rather than through a common central switch system. But as has been observed by [15] a common switch, with its own set of rules for participation, technical and operational issues, improves coordination and customer experience, and allows for a much faster implementation of interoperability, as compared to private switches or bilateral agreements.

Like Kenya and the other East African countries, Rwanda has an equally mature and highly competitive mobile money landscape [16]. Again similar product offerings are on offer by the different mobile money providers and these include balance maintenance, deposits, withdrawals and transfer of funds with convenience that is not currently being met by the commercial banks to the poor unbanked.

Despite mobile money services having been operational for a long time now, Rwanda equally does not have a formalized central clearing and settlement system that offers interoperability for the mobile money providers. This study [16] reviewed the regulation of mobile money aspects in Rwanda and considered among others, interoperability for the country with the aim of fostering a conducive financially inclusive society. The study proposes a light handed regulatory approach owing to the highly technical and capital intensive nature of the mobile money industry.

While countries like Kenya have bilateral based interoperability models in place of a central integrator mode, Rwanda has yet been to establish one. New regulation in Rwanda requires interoperability of all payment systems before integration could be realized. What has rather been observed in this market however, is the fact that subscribers transacting across networks through the use of agents. For example, an MTN user can always send money to a Tigo user, but the receiver will have to visit an MTN agent to withdraw the cash and the charges are slightly higher. In addition, if the subscriber then wants to use that cash on the Tigo system, he will have to visit a Tigo agent to make the deposit – so getting cash from a deposit in one system to a deposit in another requires visiting two agents.

Interoperability between the Rwanda banking system and mobile money services is similarly available in a weak form – it requires a physical visit to a bank branch. The next step in interoperability would allow the remote payment from an account on one provider directly into the account of another via a command from a mobile phone or bank branch.

There are four different mobile network operators all providing mobile money services to their subscribers in Tanzania [17]. Tanzania is one of the most successful mobile money markets in the world with more than 25% of the population being active mobile money users (with almost 11 million in December 2013) and transacting an estimated USD 2 billion in transactions per month in 2014 [17].

According to a study [18] by the GSMA on account to account (A2A) interoperability models in Tanzania and Pakistan, A2A interoperability was launched in Tanzania in 2014, and in Pakistan in 2015. The study found that in both Pakistan and Tanzania, the regulatory environments were enabling for A2A interoperability and that providers freely choose the technical model that best suited their commercial interests rather than being restricted to a pre-determined or preferred model defined by regulation. This has led to Tanzanian mobile money providers opting for bilateral point to point integrations as a preferred model for interoperability.

As been pointed out [1], bilateral models may seem easy to deploy where there are limited parties involved but later suffer several disadvantages including the increase in complexity with number of parties, duplication of efforts and an increase in complexity of maintenance over time.

Literature studied showed a number of different approaches to interoperability employed in different countries. One such an approach is the use of a Central Bank led national switching system for clearing and settlement.

Mobile money services in Zambia are regulated by the Central bank and therefore, this makes the use of a central switch an ideal and suitable enough approach to interoperability. So far, the technological setup used in such an approach has been with a central database system.

A number of problems with this approach have pointed out including, complexity of integration, introduction of a single point of failure and lack of trust. This paper therefore proposed a blockchain based solution approach to address these shortfalls.

A number of blockchain use cases are presented in the following section to highlight some of the properties of blockchain that make it a suitable technology to address these problems.

B. Blockchain use Cases

The A number of blockchain based solutions have been proposed by various researchers across different industries over the last few years that the technology has matured. This section highlights some of these solutions.

Firstly, [19] examined the use of Distributed Ledger Technologies (DLT) in the area of payments, clearing and settlement and identified both a number of opportunities and challenges facing its long-term implementation and adoption. Further calls for tamper-resistant data stores solutions are made in [20] by proposing the use of a write once and read multiple times data storage solution.

Similar calls are echoed in an attempt to solve problems in the management of clinical records [21]. It is argued that a blockchain technology has the potential to solve the records management problems by providing a single, secure, decentralized storehouse of clinical data for all patients.

A solution for parking slot management in a trust less network is proposed here [22] which seek to provide a platform capable of being used without a third trusted party.

In [23], a decentralized traceability system based on Internet of Things (IoT) and blockchain is proposed for the food industry. While [24] proposes a hybrid architecture for supply chain management based on a set of private distributed ledgers for storing sensitive customer information and a public ledger where a hash of each private event is stored along with the monitoring events.

Like many such similar use cases proposed, the goal to implement a secure and trusted system that takes advantage of the blockchain properties of transparency, immutability and shared consensus [25].

IV. Research Process

This study was guided by three (3) main objectives. Firstly, a targeted survey and interviews were conducted to establish how mobile financial services are currently implemented in Zambia. Further, literature and documentation on mobile money system and service implementation was consulted to understand how they are setup. The goal was to try to establish and highlight short falls and inefficiencies in implementation that prevent interoperability and thereby identify opportunities for improvements.

Secondly, an analysis as to whether a conceptual model for inter operator mobile financial transactions payments, clearing and settlement in a secure, transparent and trusted manner could be proposed and designed. The goal was to establish if blockchain technology would be an ideal technology to achieve the proposed design.

Finally, we carried out an implementation of a prototype that demonstrates Blockchain security services in a permissioned and regulated environment. The designed system was a prototype system in which amounts being interchanged between mobile money providers are managed as assets on a permissioned blockchain. The system runs a distributed shared ledger which prevents amount theft as well as fraud such as transferring invalid amounts, or transferring multiple copies of an amount, by leveraging the consistency features of the blockchain.

A. Survey Design

A list of interview questions were designed into a survey and administered to a target audience of respondents, deliberately selected according to set criteria. Further, walk in interviews were conducted with subject matter experts to validate and verify researched literature and documentation on mobile money systems and service implementation. The goal was to try to establish and highlight short falls and inefficiencies in implementation that prevent interoperability and thereby identify opportunities for improvements in the solution design.

B. Survey Participant Target Group

The research participants were purposively selected basing on their expertise, experience and skills relating to the subject under study in order to get rich and relevant information. Survey participation was drawn from employees of Zambia's mobile money operators and employees from Zambia's mobile money regulatory and supervisory authority, the Bank of Zambia. The operators included the major Mobile Network Operators (MNOs), Airtel (Airtel Money), MTN (MTN Mobile Money) and Zamtel (Zamtel Kwacha).

Participation was further extended to non-MNO providers who have been running money transfer services on mobile and have since extended their product offerings to include the mobile wallet feature on their services, which allows customers to hold value and transact off those accounts. These included Zoona (who run the Zoona Plus wallet), Broadpay (who run the Broadpay wallet) and cGrate (who run the Konse Konse wallet).

C. Survey Sampling Rationale

Due to the specialized nature of the data that the research required, survey respondents had to be conveniently sampled. The Bank of Zambia, for example, is the regulatory authority that supervises and regulatory financial services providers in Zambia. They do this through among others registration and designation of payment systems and institution as well as oversight of both systemic and non-systemic payment systems.

The central bank is also responsible for the clearing and settlement infrastructure and processes in the country. It was felt strongly therefore, that they would be well positioned to provide information on payment system interoperability from regulatory and standards perspectives. Participation therefore, was also drawn from a number of Bank of Zambia staff with varying specializations. These included Payments Systems specialists, Financial Institutions Supervision specialists, Information Systems specialists and Information Systems Security specialists.

D. Blockchain Decision Model

For the second part of the study, we looked at whether and how a blockchain based solution would be ideal for this use case. This was necessary because unlike in Bitcoin's permission-less blockchain, where any writer and reader can join at any time, permissioned blockchains have restricted read and write access thus share close similarities with a centralized database systems. This thus naturally brings up the question whether a blockchain is better suited than a centralized database.

A flow chart based decision model was therefore, adopted and used to determine the suitability of the technology to be adopted as proposed by Wüst and Gervais [26]. The model used here is shown in Fig. 1. Other such similar models have been proposed [27], [28]. This model was found more suitable as it provides a detailed description of the decisions leaving less room for misinterpretation. The model consists of a decision tree based on the following scenario properties:

1) Storing state: Refers to the need of storing data that may change both in volume and in content over time.

2) Existence of multiples writers: These are the writers that have a common interest in agreeing on the validity of the stored state.

3) Need for trusted third party: A Trusted Third Party (TTP) is a centralized entity that could manage changes and updates the state. A TTP, if present, may also control who can read the state stored.

4) Are all writers known: This refers to knowing the identity of all writers.

5) Are all writers trusted: When writers are trusted, they are expected not to behave maliciously. When writers are not trusted, they may behave maliciously.

6) Public verifiability of sate: This property determines who may read the state stored on the blockchain, and verify the integrity of the ledger.

Based on these six properties, the model determines one of four possible solutions as the best solution for the scenario:

- Permissionless blockchain: Anyone may join the network and read from the state stored, and write to the blockchain.
- Public permissioned blockchain: A limited set of participants may write to the blockchain. Anyone may join the network and read the state.
- Private permissioned blockchain: A limited set of participants may join the network, and write a new state. Only this set can read the state.
- Don't use blockchain: This end state is reached when one of the properties (1), (2), (3), or (5) above is not met.

E. Proposed Solution Design

A formal software development methodology was followed in the design and implementation of the solution prototype proposed. Object Oriented Analysis and Design methodology using the Object Modeling Techniques (OMT) phases to model the different aspects of the prototype was used. The proposed framework consists of a common replicated ledger in which transferred amounts are managed as assets on a permissioned blockchain based on Hyperledger Fabric [29] as summarized in Fig. 2.

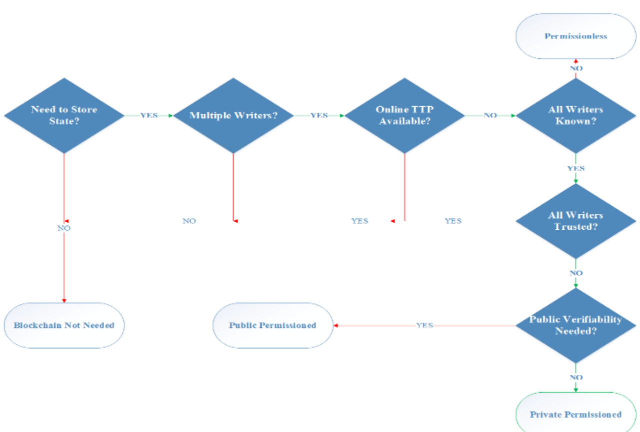


Fig. 1. Decision Model Adopted on Blockchain use Case (Source: wüst and Gervais [26]).

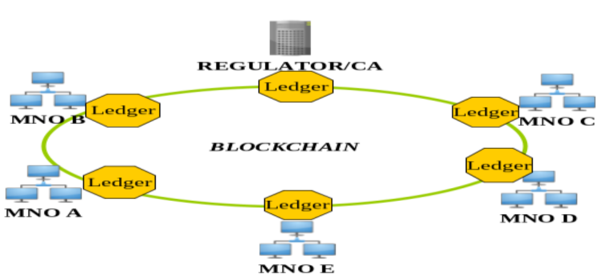


Fig. 2. Proposed Highlevel Network Architecture.

Hyperledger Fabric is an open source permissioned distributed ledger technology (DLT) platform, designed for use in enterprise contexts [30]. It is designed for business use cases where the blockchain is operated by a set of known, identified, and often vetted participants. This capability is known as a permissioned blockchain. A permissioned blockchain provides a way to secure the interactions among a group of entities that know each other, have common business interests, and want to manage a decentralized network (rather than turning management of their ledgers over to a single party) [31].

By relying on the identities of the peers, a permissioned blockchain can use traditional crash fault tolerant (CFT) or Byzantine fault tolerant (BFT) consensus protocols that are used by many other distributed programs [32]. Hyperledger Fabric offers high levels of performance, protection, and transaction privacy.

Fabric was chosen because of its highly modular and configurable architecture that makes it adaptable to a number of use cases. Fabric also supports the use of general purpose programming languages such as Java in the development of smart contracts and therefore, was an ideal choice for this prototype.

Blockchain approach was used to provide key security requirements of confidentiality, origin authentication, non-repudiation and availability.

1) Blockchain network architecture: The network layout is depicted in Fig. 2 as a shared, replicated, permissioned distributed ledger where all participants have a copy of the ledger alongside their data. The blockchain architecture gives participants the ability to share a ledger that is updated every time a transaction occurs through peer-to-peer replication.

The Fabric network consists of the following basic components [33]:

1) Ledger: A ledger which consists of the world state and the blockchain. The world state contains the status of all assets that are tracked on the ledger (who owns a particular asset, for example), while the blockchain contains a history of all state changes. Ledgers are replicated across a channel and stored on peers.

2) Peers: These are the transaction endpoints for organizations and make up much of the physical structure of a network. They are maintained by members (organizations) whose identities are known by the blockchain network. Peers can maintain multiple ledgers (they have one for every channel they are a member of) and endorse transactions.

3) A channel: This contains a subset of network members who want to communicate and transact privately. Ledgers are channel specific (that is, every channel has a separate ledger). Only the peers on a channel can see the assets and transactions for its ledger. As a result, channels ensure privacy for participants within the network.

4) Chaincode: Hyperledger Fabric smart contracts are implemented in chaincode. When an application needs to interact with the ledger, it invokes these contracts by sending transactions into the Fabric network. This is the case because chaincode predominately interacts only with the database component of the ledger and not the historical transaction log.

5) Orderer: The Ordering Service, usually composed of multiple orderers, provides consensus and ordering of transaction. It does so by bundling transactions into blocks, which are then added to the blockchain.

6) The Certificate Authority (CA): This identifies all entities in the network: Peers, the ordering service, and the participants who are submitting transactions and accessing the ledger. These identities are provided and secured by using a public key infrastructure (PKI). Peers use the CA to cryptographically sign transactions and contracts, whereas participants use the CA to prove that they have a right to access the network.

7) SDK: The Hyperledger Fabric Client SDKs enable interaction between your client application and your blockchain network. With support for multiple languages, the SDK contains APIs that allow an app to connect to and to access the smart contracts and the ledger for the channel the peer is on.

Fig. 3 shows the main nodes and components that make the proposed solution. Each participants (labelled as MNO in Fig. 2) maintains their own mobile money systems. As part of the Fabric network, each participant also runs peers which allows them to connect to the rest of the blockchain network. These peers receive transaction requests from participant systems though an Application Programming Interface (API) provided by the Software Development Kit (SDK).

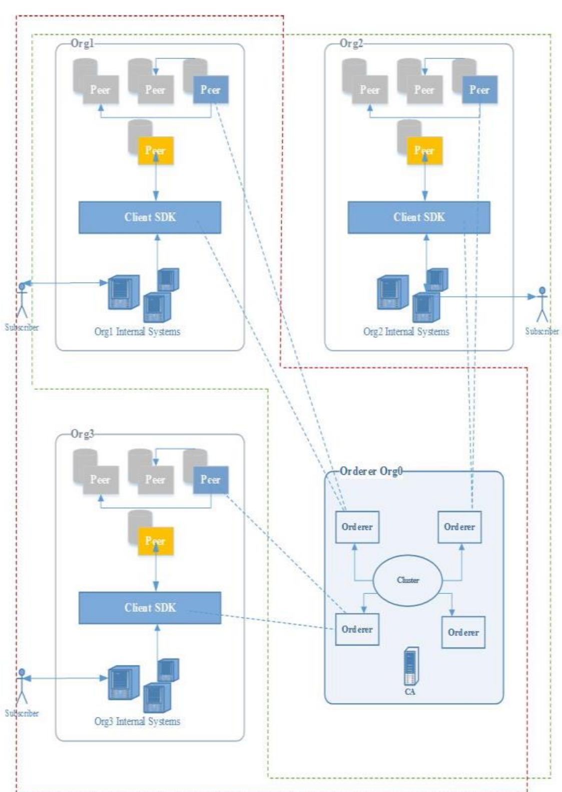


Fig. 3. Blockchain Solution Architecture.

Each pair of participants (Org1 and Org2 B for example) connect through a separate channel interface that allows them to maintain data privacy between the two. The Orderer node is responsible for ordering and writing transaction requests to the ledger before replication.

2) Use Case Model: The main asset that is transacted on the proposed network is a transfer and this represents a request made by one subscriber through a participant to transfer an amount to another subscriber on a different participant's network. Fig. 4 shows the main use cases in the system while Fig. 6 shows the states through which the transfer transitions.

Two main classes of actors are identified in the ecosystem and these are the direct participants and non-direct participants. The direct participants are the mobile money providers that directly take part on the blockchain and the clearing house which is a special institution (the "settler") responsible for netting and settlement. The non-direct actor is the subscriber who participates through the Operator and represents the mobile money subscribers.

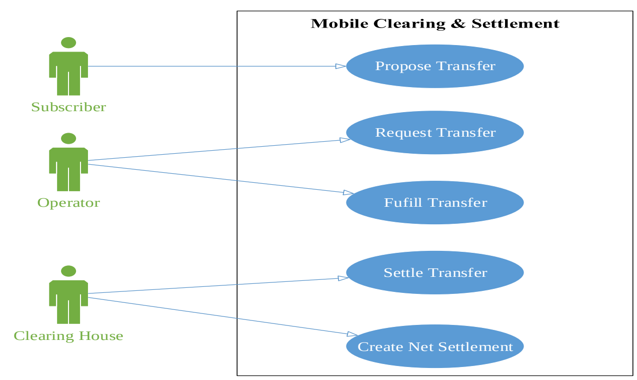


Fig. 4. Solution use Case Diagram.

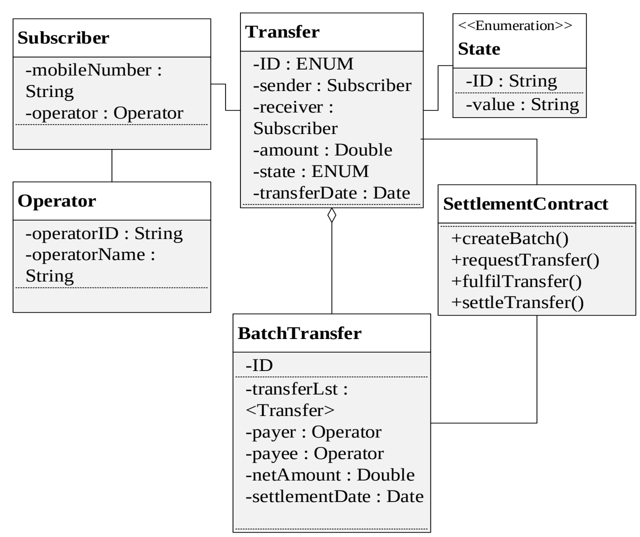


Fig. 5. Solution Class Diagram.

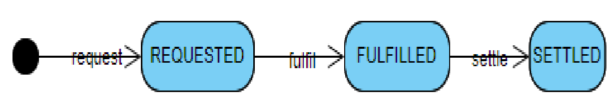


Fig. 6. State Transition Diagram for the Transfer Asset.

A special program called a smart contract was implemented that models this transaction logic that transitions the transfer between their different states. Smart contracts allowed us to define the key business processes and data that are shared across the different organizations collaborating in the network. Fig. 5 shows the class model that captures the smart contract and depicts the main objects that make up the smart contract.

V. Results and Analysis

In this section we present the results of the study. Key findings of the baseline are presented and their subsequent application to the study. Highlights of the prototype system designed are also given in terms of code artefacts as well as screenshots of the experimental Fabric network that was setup.

A. Baseline Study Results

Participation was drawn from providers with varied subscriber bases (Fig. 7) and each using different platforms and reported facing integration challenges (Fig. 8) Overall, on inter operator integration, the general feeling was that it was manageable and could be eased with the use of a central integrator rather than having every operator to integrate individually with every other provider.

B. Decision Model Analysis

A Decision model was adopted and used to assess blockchain suitability to this use case [26]. Key findings from the baseline study were used in the flow chart decision tree as prescribed in this model and it was established that for this particular use case, we could make use of a permissioned blockchain as a technology. Table I summarizes these findings.

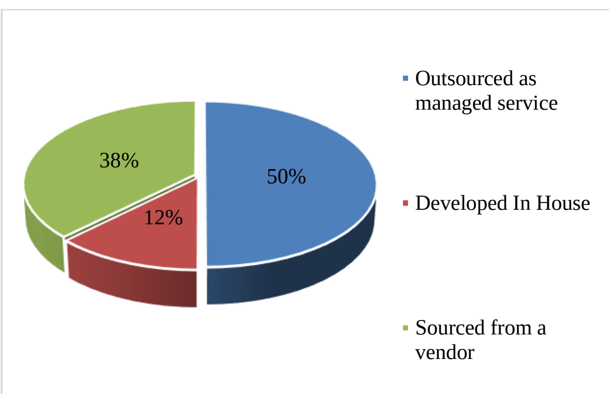


Fig. 7. Mobile Money Operator's Platform Sources.

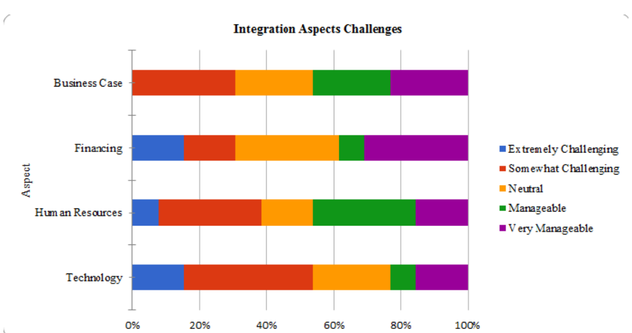


Fig. 8. State Transition Diagram for the Transfer Asset.

TABLE. I. Key Survey Findings

Decision Model Analysis		
Decision State	***Finding Description***	***Result***
Storing state	Existence of different independent mobile money operators	YES
Existence of writers	Existence of technological platforms or systems on which these operators run their services	YES
Trusted Online Third Party	Controlled access to the network with permissioning.	NO
Are all writers known	The need for integration among these systems to provide interoperability	YES
Are all writers trusted	Security and privacy of transactions	NO
Public verifiability of state	Security and privacy of transactions	NO

a. Decision tree based on [26]

A number of important aspects such as the need to store state, existence of multiple writers were used to arrive at the decision of the solution. Other aspects like the need for central management and the relatively low number of writers were also considered to arrive at the decision.

C. Chaincode Implementation

This section highlights the main implementation aspects of the prototype system as proposed. Presented are the code snippets of the major parts of the chaincode that drives the smart contract on the fabric network.

1) The contract class: The main smart contract classes is the SettlementContract class and this contains the transaction definitions for the system. These are the request, fulfil, settle and batch transactions that have been defined and which move the assets through the application life cycle (Fig. 6).

The SettlementContract class implements the ContractInterface and so the Settlement contract uses built-in features of these classes, such as automatic method invocation, a per-transaction context, transaction handlers, and class-shared state. Fig. 9 shows code snippet of this implementation detail.

This class contains implementation of a number of methods that control application lifecycle. Firstly, the requestTransfer (Fig. 10) method creates a new transfer context object between two participants (sender and receiver) which is saved on the ledger as an asset.

The fulfilTransfer (Fig. 11) method is another transaction method and it transitions a transfer object in REQUESTED state and sets it to the FULFILLED state (after the receiver has fulfilled the transaction as confirmation that funds have been moved that participant's account).

The settleTransfer (Fig. 12) method is also another transaction method and it transitions a transfer object in FULFILLED state and sets it to the SETTLED state. This method is called by the createBatch method during the net settlement process at the end of business day.

Finally, the createBatch method is a settlement process method that is called to collate all transfers between any pair of participants and create a BatchTransfer asset which is stored on the ledger.

This class and methods make up the transaction logic part of the system and represent the control flow logic of processing. Next we highlight the object implementation which represents the main assets.

```java
@Contract(name = "org.momo-switch.transfer", info = @Info(title = "Momo Contract", description = "",
@Default
public class SettlementContract implements ContractInterface {
```

Fig. 9. Settlement Contract Class Definition.

```java
@Transaction
public Transfer fulfilTransfer(TransferContext ctx, String receiver, String transferNumber) {

    // Retrieve the current transfer using key fields provided
    String transferKey = State.makeKey(new String[] { transferNumber });
    Transfer transferToBeFulfilled = ctx.transferList.getTransfer(transferKey);

    // Check transfer is indeed in REQUESTED state
    if (transferToBeFulfilled.isRequested()) {
        transferToBeFulfilled.setFulfilled();
    } else {
        throw new RuntimeException(
                "Transfer " + transferNumber + " already fulfilled or is not requested");
    }

    // Update the transfer state on the ledger
    ctx.transferList.updateTransfer(transferToBeFulfilled);
    return transferToBeFulfilled;
}
```

Fig. 10. Request Transfer Method.

```java
@Transaction
public Transfer requestTransfer(TransferContext ctx, String sender, String receiver, String transferID, String transferDateTime,
        String settlementDateTime, int amount) {

    // create an instance of the transfer
    Transfer transfer = Transfer.createInstance(sender, receiver, transferID, transferDateTime, settlementDateTime,
            amount,"");

    // Smart contract, rather than transfer, moves transfer into REQUESTED state
    transfer.setRequested();

    // Add the transfer to the list of all transfers in the ledger
    ctx.transferList.addTransfer(transfer);

    // Return created transfer to caller of smart contract
    return transfer;
}
```

Fig. 11. Fulfil Transfer Method.

```java
@Transaction
public Transfer settleTransfer(TransferContext ctx, String issuer, String transferNumber) {

    String transferKey = Transfer.makeKey(new String[] { transferNumber });

    Transfer transferToBeSettled = ctx.transferList.getTransfer(transferKey);

    // Check transfer is not already SETTLED
    if (transferToBeSettled.isFulfilled()) {
        transferToBeSettled.setSettled();
    } else {
        throw new RuntimeException(
                "Transfer " + transferNumber + " is not ready for settlement ");
    }

    // Update the transfer
    ctx.transferList.updateTransfer(transferToBeSettled);
    return transferToBeSettled;
}
```

Fig. 12. Settle Transfer Method.

2) *The main object classes:* The main object classes that represent assets on the ledger are the Transfer and the BatchTransfer classes. These classes have member valuables that represent the properties of the assets and have respective createInstance methods which are used to initialize their respective objects so as ensure instantiation of these objects is through a transaction rather than through the classes.

These classes also extend the State class which is used to control lifecycle states of the assets and represents the ledger level Fabric state database. Fig. 13 shows the main parts of these classes.

The other object classes include the Operator and the Subscriber and these used to represent logical member variables for the respective objects for easier management. Code snippets showing implementation are presented in the appendix for those.

```java
@DataType()
public class Transfer {

    //Transfer State values
    public static final String REQUESTED = "REQUESTED";
    public static final String FULFILLED = "FULFILLED";
    public static final String SETTLED = "SETTLED";
    public static final String REJECTED = "REJECTED";

    @Property()
    private Subscriber sender;
    @Property()
    private Subscriber receiver;
    @Property()
    private Double amount;
    @Property()
    private String transferDate;
    @Property()
    private String transferID;
    @Property()
    private String state = "";
```

```java
@DataType()
public class BatchTransfer {
    private String ID;
    private ArrayList<Transfer> transferList;
    private Operator payer;
    private Operator payee;
    private Double netAmount;
    private String settlementDate;


}
```

Fig. 13. Asset Class Definitions.

VI. CONCLUSION

The study proposed the use of blockchain technology to solve the problem of mobile money interoperability in Zambia. A structured approach was used to confirm the gap and then decide a technological solution through the use of a structured decision model for careful determination. We further designed a prototype system on the Hyperledger Fabric network which could develop in an Object Oriented language such as Java for deployment.

We can thus conclude that mobile money interoperability settlement is a valid use case for a permissioned blockchain technology and would be an ideal solution approach rather than the traditional central processing database systems.

VII. LIMITATIONS AND FUTURE WORKS

This study focused on a gap verification of the interoperability problem as well as a technical implementation of a prototype solution. The prototype also only considered the funds transfer between participating entities and their subsequent settlement and did not look at other technical aspects such as the regulatory aspects and the financial and business sides of the ecosystem. The prototype was only experimental and could only be deployed on a development network and not in a live network with integration with mobile network operators for a more real world demonstration.

ACKNOWLEDGMENT

Acknowledgment of special thank you goes to the University of Zambia (UNZA) Department Of Computer Science and Department of Electronics and Electrical Engineering for the support and guidance during the study.

REFERENCES

[1] D. Clark and G. Camner, "A2A Interoperability: Making mobile money schemes interoperate," no. February, 2014.

[2] M. Tompkins, A. Olivares, and M. Tompkins, "Clearing and Settlement Systems from Around the World : A Qualitative Analysis Clearing and Settlement Systems from Around the World : A Qualitative Analysis by," 2016.

[3] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, "An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends," Proc. - 2017 IEEE 6th Int. Congr. Big Data, BigData Congr. 2017, no. June, pp. 557–564, 2017.

[4] Bank of Zambia, "Payment Systems Vision and Strategy 2018-2022," Bank of Zambia, Lusaka.

[5] T. TEMBO, "National financial switch to start this year," Daily Mail, 2017. [Online]. Available: http://www.daily-mail.co.zm/national-financial-switch-to-start-this-year/. [Accessed: 21-Jan-2018].

[6] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," pp. 1–9, 2013.

[7] D. C. Mills et al., "Distributed Ledger Technology in Payments, Clearing, and Settlement," 2016.

[8] M. Bourreau and T. Valletti, "Competition and Interoperability in Mobile Money Platform Markets: What Works and What Doesn't? (*)," Commun. Strateg., 2015.

[9] R. B. of I. Department of Payment and Settlement Systems, "Prepaid Payment Instruments (PPIs) – Operational Guidelines for Interoperability," no. 808. Mumbai, India, pp. 2–4, 2018.

[10] D. Kumar, T. A. Gonsalves, A. Jhunjhunwala, and G. Raina, "Mobile payment architectures for India," Proc. 16th Natl. Conf. Commun. NCC 2010, no. 1, pp. 4–8, 2010.

[11] K. Kumar and M. Tarazi, "Interoperability in Branchless Banking and Mobile Money. Consultative Group to Assist the Poor (CGAP).," CGAP, 2012. [Online]. Available: http://www.cgap.org/blog/interoperability-branchless-banking-and-mobile-money-0.

[12] M. Mamabolo, "Kenya's Central Bank gives mobile money interoperability thumbs up," IT WebAfrica, 2018. [Online]. Available: http://www.itwebafrica.com/fintech/842-kenya/243988-kenyas-central-bank-gives-mobile-money-interoperability-thumbs-up. [Accessed: 20-Dec-2018].

[13] Central Bank of Kenya, "Mobile Money Interoperability," no. January, p. 2018, 2018.

[14] M. Mamabolo, "Kenya pilots mobile money interoperability," IT WebAfrica, 2018. [Online]. Available: http://www.itwebafrica.com/mobilex/309-kenya/242443-kenya-launches-mobile-money-interoperability-pilot-today. [Accessed: 20-Dec-2018].

[15] M. Bourreau and S. Hoernig, "Interoperability of Mobile Money: International Experience and Recommendations for Mozambique," no. December, 2016.

[16] J. Argent and J. A. Hanson, "The Regulation of Mobile Money in Rwanda," J. ICT, 2013.

[17] Bank of Tanzania, "National Payment System Directorate Statistics:" [Online]. Available: http://www.bot-tz.org/PaymentSystem/statistics.asp. [Accessed: 18-Dec-2018].

[18] GSMA, "Choosing a Technical Model for A2A Interoperability: Lessons from Tanzania and Pakistan," no. December, 2015.

[19] D. C. Mills et al., "Distributed Ledger Technology in Payments, Clearing, and Settlement," Ssrn, 2016.

[20] B. Smith and K. Christidis, "IBM Blockchain: An Enterprise Deployment of a Distributed Consensus-based Transaction Log," Proc. Fourth Int. IBM Cloud Acad. Conf., pp. 1–4, 2016.

[21] D. Ivan, "Moving Toward a Blockchain-based Method for the Secure Storage of Patient Records," 2016.

[22] R. J. Reisman, "Air Traffic Management Blockchain Infrastructure for Security , Authentication , and Privacy," pp. 1–14, 2019.

[23] F. Tian, "A Supply Chain Traceability System for Food Safety Based on HACCP, Blockchain & Internet of Things," 14th Int. Conf. Serv. Syst. Serv. Manag., vol. 323, no. 2, pp. 511–519, 2015.

[24] H. Wu, Z. Li, B. King, Z. Ben Miled, J. Wassick, and J. Tazelaar, "A distributed ledger for supply chain physical distribution visibility," Inf., vol. 8, no. 4, pp. 1–18, 2017.

[25] A. Baliga, "The Blockchain Landscape Office of the CTO," Persistent Syst. Ltd, p. 21, 2016.

[26] K. Wüst and A. Gervais, "Do you need a Blockchain?," IACR Cryptol. ePrint Arch., no. i, p. 375, 2017.

[27] M. Peck, "Do You Need a Blockchain?," IEEE Spectrum, 2018. [Online]. Available: https://spectrum.ieee.org/computing/networks/do-you-need-a-blockchain. [Accessed: 12-Dec-2018].

[28] T. Koens and E. Poll, "What Blockchain Alternative Do You Need ?".

[29] E. Androulaki et al., "Hyperledger fabric," pp. 1–15, 2018.

[30] Varun Raj, "Hyperledger Fabric Architecture: Explained in detail," 2018.

[31] J. Garzik, "Public versus Private Blockchains. Part 1: Permissioned Blockchains," 2015.

[32] M. Sethumadhavan, "On Blockchain Applications: Hyperledger Fabric And Ethereum," Int. J. Pure Appl. Math., vol. 118, no. 18, pp. 2965–2970, 2018.

[33] C. Cachin, "Architecture of the Hyperledger Blockchain Fabric," Work. Distrib. Cryptocurrencies Consens. Ledgers (DCCL 2016), 2016.

Sign Language Semantic Translation System using Ontology and Deep Learning

Eman K. Elsayed[1], Doaa R. Fathy[2]
Mathematical and computer science
Al-Azhar University
Cairo, Egypt

***Abstract*—Translation and understanding sign language may be difficult for some. Therefore, this paper proposes a solution to this problem by providing an Arabic sign language translation system using ontology and deep learning techniques. That is to interpret user's signs to different meanings. This paper implemented ontology on the sign language domain to solve some sign language challenges. In this first version, simple static signs composed of Arabic alphabets and some Arabic words started to translate. Deep Convolution Neural Network (CNN) architecture was trained and tested on a pre-made Arabic sign language dataset and on a dataset collected in this paper to obtain better accuracy in recognition. Experimental results show that according to the pre-made Arabic sign language dataset the classification accuracy of the training set (80% of the dataset) was 98.06% and recognition accuracy of the testing set (20% of the dataset) was 88.87%. According to the collected dataset, the classification accuracy of the training set was 98.6% and Semantic recognition accuracy of the testing set was 94.31%.**

Keywords—Deep Learning (DL); ontology; sign language translation

I. Introduction

Sign language considered the only way for communication between deaf, hearing-impaired and normal people. According to the World Health Organization (WHO), there are about 466 million people around the world have disabling hearing loss, and more than 28 million of these are Americans, 13 million people within Egypt across all age groups [1]. People use sign language gestures as a means of non-verbal speech to express their thoughts and emotions.

Sign language has two types of gestures: static gestures and dynamic gestures [2]. Due to Arabic language variations (delicates), Arabic Sign Language has multiple country variants and dialects. It differs from one country to another and sometimes differs within the same country. Deaf and hearing-impaired people have all the right to speak to each other even for different sign languages. Sign languages need an intelligent system to translate sign language to another based on natural languages. It is hard for most people who are not interested in sign language to communicate without an interpreter. Automatic sign language translation systems are important for solving these problems.

Sign Language Recognition Systems can be developed by using either sensor-based technique or image-based technique [3]. Many types of research focused on sign language translation to text and spoken language and vice versa. But this is not easy to be done by machines since it depends on the natural language processing and image recognition. On the other hand, the traditional way of translation needs a translator that specializes in sign language may not be present in all situations, especially when the deaf continues with the person hearer. Translation of Arabic sign language (ArSL) is facing many challenges; the lack of linguistic studies on ArSL, sign language is assumed to be a universal language.

Deep Learning [4] algorithms have been successfully applied to image recognition problems. Deep learning involves neural networks with more than one hidden layer, has been used successfully in face recognition, speech recognition and natural language processing problems [5, 6]. Deep learning successfully implemented for human gesture recognition in recent years.

Ontology is a formal representation of knowledge; it handles a set of concepts in a specific domain, their relationships, and properties [7]. WordNet is a linguistic resource containing not only words of the targeted language but also synsets and semantic conceptual relations between them. These relations provide semantic information about concepts and their original words. These concepts and their relations are exploited to improve Information Retrieval, Text Classification, and Text Summarization Technically. Arabic WordNet lexicon provides a good semantic structure for computing the semantic similarity between words.

The contribution of this paper is: To enhance sign language translation using the power of semantic web technologies (i.e. ontologies) and deep learning. Ontology used in building multi-sign language ontology (MSLO) and deep CNNs used in Arabic sign language feature extraction and recognition processes. As a case study, the proposed method applied in Arabic sign language semantic translation system.

The rest of this paper is organized as follows; the next section presents some related works. Then the proposed method will present in Section III. Subsection A in Section III presents the proposed Multi Sign Language Ontology (MSLO) and subsection B presents the details of the proposed "Semantic Deep Learning". Section IV presents the application of the proposed method on a case study in Arabic Sign Language, while Section V presents the analysis of results. Finally, conclusion and future works.

II. Related Works

In recent years, several research projects in Arabic static sign language translation systems have been developed. Author in [8] presented an automatic visual technique that translates individual Arabic sign to text word. Geometric features of the hands are employed to formulate the feature vector. Euclidean distance classifier is applied for the classification stage. A dataset of 30 isolated words of the hearing-impaired children was developed. The system has a recognition rate of 97%.

The SVM method and HOG descriptors were used to recognize Arabic Sign Language alphabets [9]. This system extracts the HOG descriptor and transfer features to One Versus All soft-margin (SVM). The proposed system has reached an accuracy of 63.5% for Arab Alphabet signs

Hidden Markov Model (HMM) technique successfully recognizes the alphabet and numbers signs of Arabic sign language to text or speech [10]. The proposed algorithm uses hand geometry and the different shape of a hand in each sign for classifying letters shape by using HMM.

Also, there exist many reported research systems related to Arabic sign language recognition based on deep learning. Authors in [11] proposed a new system based on the CNNs, fed with a real dataset, this system recognized automatically numbers and letters of Arabic sign language. To validate the system, they done a comparative study that shows the effectiveness and robustness of their proposed method compared to traditional approaches based on k-nearest neighbors (KNN) and support vector machines (SVM).

ElBadawy et al. [12] developed the system for Arabic sign language recognition. In this system, CNN was used to recognize 25 gestures from the Arabic sign language dictionary. The system achieved 98% accuracy for observed data and 85% average accuracy for new data.

Although a variety of methods have been proposed in recent years to recognize static hand gestures, most of them worked only on translating text to sign or sign to text. They did not take care of the semantics of the translated text. In another side, some researches work on sign language translation using Ontology. Author in [13] presents a proposed system for semantically translating Arabic text to Arabic Sign-Writing in the jurisprudence of prayer domain using ontology as a semantic web technology. The system designed to translate Arabic text by applying Arabic Sign Language (ArSL) grammatical rules as well as semantically looking up the words in Ontology. Author in [14] focuses on the development of information technology for Ukrainian Sign Language translation based on grammatically augmented ontologies.

III. The Proposed Ontology and Deep Learning Method

The proposed method is divided into two main parts: building ontology-based and combining ontology with deep learning.

A. *Multi Sign Language Ontology (MSLO)*

The proposed Multi Sign language Ontology (MSLO) is linguistic Ontology takes the advantages of WordNet and multilingual WordNet on sign language domain. But analysis of the nature of sign language leads us to create some different relations to solve some of sign language challenges.

MSLO extends the WordNet relations to deal with sign language. The sign language synsets are created in correspondence with the WordNet synsets, and semantic relations are imported from the corresponding synsets. It assumed that if there are two synsets and relations are holding between them, the same relations hold between the corresponding synsets in sign language. Generally, sign express word meaning (concept), it could be said that it is important to sign a concept rather than a word. MSLO solves some sign language challenges. The following three examples of these challenges:

1) Each sign means a bag of words in each language (synonyms). For example, Fig. 1 presents the sign of “car" in English language. The same sign could represent the words: Vehicle, automobile, transportation, and auto. The same sign also used to represent the words "سيارة" , "مركبة","حافلة","عربة" in the Arabic language.

2) The same meaning has different signs and vice versa. In MSLO, the two words related semantically to each other may have different signs. For example, the word “bad” in English has the sign (a) in Fig 2. The word “سيئ” in Arabic has the sign (b) in Fig. 2.

Also, one sign may use to represent different words such as; sign in Fig. 3, represents the word "لا", the Character "ب" and number "1" in the Arabic language. This challenge solved in MSLO ontology, where sign gesture interprets based on its class and domain.

Fig. 1. The Word “CAR” in Sign Language.

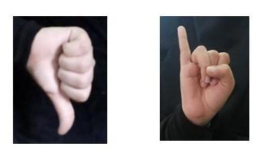


Fig. 2. (a)"سيئ" Gesture Sign (b) Bad Gesture Sign.

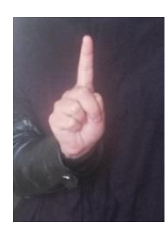


Fig. 3. Gesture Sign of Word "لا", Character "ب" and Number "1".

3) Each person has his-self way to sign any word in any language. MSLO takes into consideration personality signs.

Relations were created in MSLO to solve these sign language challenges. Properties used to describe a class/concept or individuals are:

a) Object properties used for the relationship between words:

- Arabic_ meaning
- English_ meaning
- French_ meaning
- Has_the_same_sign; to relate the concepts which have the same gestures signs.

b) Data property that used to describe words are:

- *DL-Label;* to assign the word to its label in the deep learning model.
- Arabic_sign
- English_sign
- French_sign

The URL address of each sign gesture image is saved as a data property value.

c) Annotation property (label) also used to add all possible meanings of the words.

Deaf, hearing impaired and normal users can interact with the ontology via interaction system to develop and update his personal information (My-Dictionary class) and its corresponding signs.

Fig. 4 (screenshot from protégé) shows Part of MSLO Ontology. The main concepts are: English_WORDNET, French_WORDNET, and Arabic_WORDNET. Each language consists of sub-classes. Each subclass is a bag of words in one domain. There is a class for personal information in sign language to represent personal data; like name, address and birth date. Also, Fig. 4 shows an example of two English words (Good and Bad) with their meanings in Arabic and French languages are related to each other in antonym relation. Their sign language gestures' are also related in antonym relation. The English word Good with its meanings in Arabic and French languages are related to each other in synonym relation. Their sign language gestures' are also related in synonym relations.

B. *Combining Ontology with Deep Learning*

The link between ontology and deep learning is being explored. Author in [15] addresses the extraction of OWL information from raw text with deep learning. Author in [16] applies deep learning to Ontology extraction, obtaining encouraging results.

Deep learning output needed to be enriched by semantic, so the proposed method considered merging ontology with deep learning. Using ontology as a semantic layer in deep learning produces Semantic deep learning (SDL).

Convolutional Neural Network (CNN) [17] is the most popular algorithm used for implementing the deep learning technique. A CNN is composed of several layers: Convolutional Layer, Pooling Layer, and Fully-Connected Layer. In this method, Ontology layer is inserted to add semantic information to deep CNN classification output.

Fig. 5 shows a diagram of CNN with 4 hidden layers. The input layer has three neurons and each hidden layer has four neurons. Hidden layers followed by the output layer, the semantic layer, and the semantic output layer.

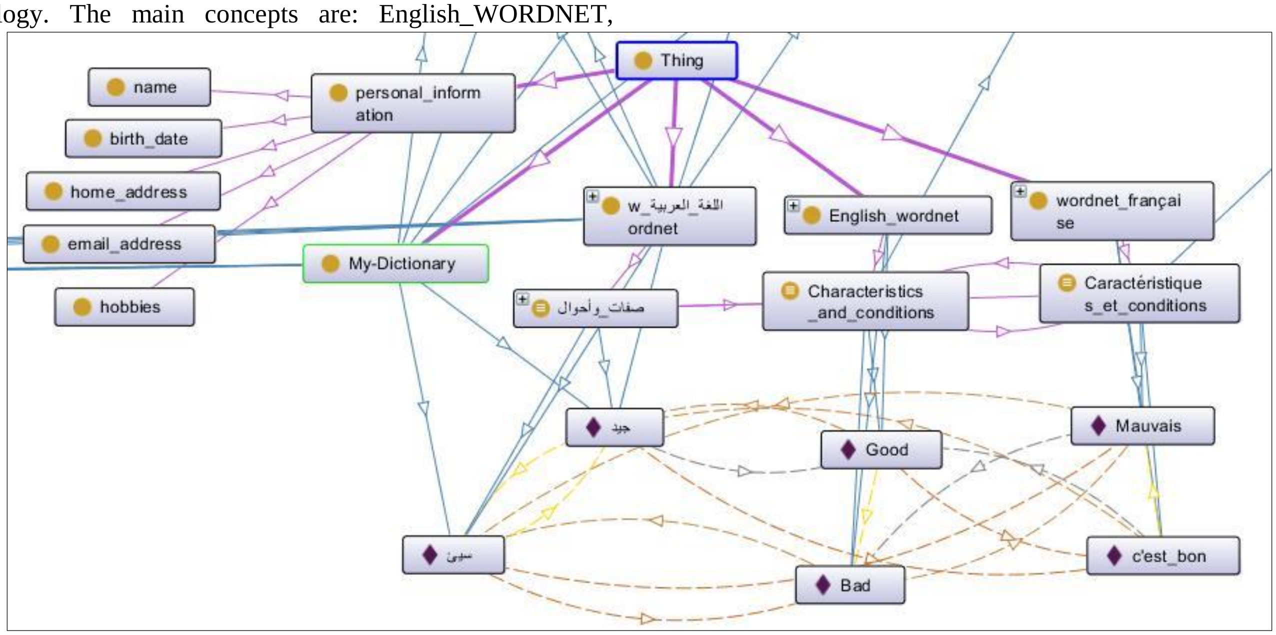


Fig. 4. Part of Proposed (MSLO).

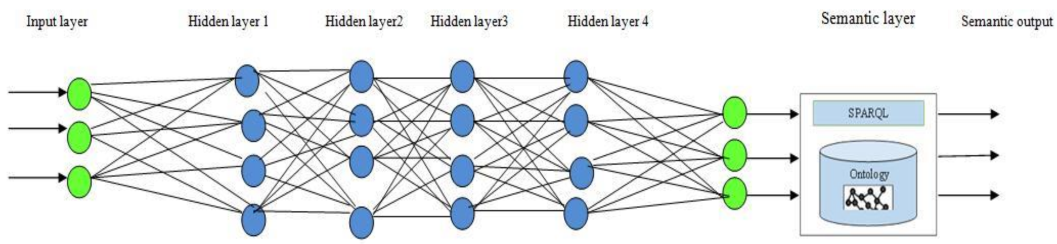


Fig. 5. Semantic Deep Learning Layers (SDL).

Semantic Deep Learning (SDL) layers are as the following:

- Convolutional layer composed of groups of neurons act as kernels. These kernels are associated with a small region of the image known as a receptive field, then convolving them (applying Convolution operation) with a specific set of weights. Where multiply each element of the filter (weights) with corresponding receptive field elements. Different sets of features extracted by sliding convolutional kernel on the image with the same set of weights. Convolution operation may categorize into different types based on the type and size of filters.

- Activation layer also called a non-linear layer; the convolution layer output is generally followed by a non-linear function called activation function to introduce non-linearity into the model. Thus, features generated by the convolution layer are transformed into another space by the activation function, and the data can be better classified. In this paper, the activation functions used are (Rectified Linear Units) ReLU.

- Batch normalization (BN) layer, is the most widely used to enhance generalization [18] and avoid over. Also, is added after a convolutional layer and its ReLU activations. Batch Normalization achieves the same accuracy with fewer training steps thus accelerating the training process.

- Dropout layer also used to avoid overfitting. The key idea of Dropout is to randomly drop units (along with their connections) during the training phase. The reduction in the number of parameters in each step of training has an effect on regularization.

- Pooling layer is used on one hand to reduce the spatial dimensions of the representation and to reduce the amount of computation done in the network. The most used pooling layer has filters of size 2 x 2 with a stride 2. This reduces the thoughts to a quarter of its original size. In this CNN max pooling is applied, which retains the maximum value within the local neighborhood of the sliding window.

- Fully connected layer, interpret these feature representations and perform high-level reasoning. Each neuron from a fully connected layer is linked to each output of the previous layer. The operations behind a convolutional layer are the same as in a fully connected layer.

- Loss layer is used to penalize the network for deviating from the expected output. This is normally the last layer of the network. Various loss function exists: softmax is used for predicting a class from several different classes, where sigmoid cross-entropy is used to predict multiple independent probabilities (from the [0, 1] interval).

- Semantic layer maps the class label to a formal representation of its meaning. When a set of words share the same fundamental meaning, they will have one label. In this layer the result of the previous layer "DL class label" used in searching Ontology to get all semantic meanings. The output of deep learning (class label) named "DL_label" represents as data property value in MSLO ontology. So each class label has its corresponding data property value to retrieve all semantic data of this class.

In this research, it was started by an image augmentation technique to adapt the training set with such images. Data augmentation has a large impact on improving the DL accuracy by reducing overfitting and increasing generalization capacity of the network [19]. Image data-augmentation increases the size of the training dataset by applying small geometric and photometric transformations.

As shown in Fig. 6, the CNN model of this paper consists of 5 convolutional layers each one followed by a max-pool layer with a filter of shape 2 x 2 with stride 2. Batch normalization and dropout layers are employed in all convolution layers. Fully-connected (FC) layer follows the convolutional layers. It has 1024 channels where is a soft-max layer to perform the classification over the number of classes. Dropout regularization and batch normalization are used also in the fully-connected layer. The dropout rates of 0.25 and 0.5 were set for the convolutional-pooling layers and the fully-connected layer, respectively.

C. Implementation

Proposed SDL method applied by Python using Keras. Tensorflow was chosen as a backend because it is mainly used in classification. Tensorflow is an open-source software library for deep learning created and maintained by Google. Anaconda used because of its flexibility to work with any operating system. Also, it has Spyder editor, a powerful python editor that has all the functionalities needed to write and run the codes. It also has the option to create an environment to install the packages. Also, Python OpenCV library used for image processing, and view the results in the 'IPYTHON' console. The MSLO implemented using Protégé as Ontology editor.

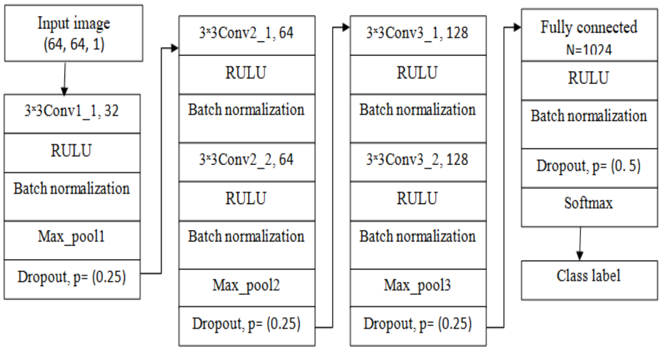


Fig. 6. Architecture of Deep Learning layers in SDL.

IV. The Case Study on Arabic Sign Language and its Semantic Translation System

A. Arabic Signs Dataset

In this paper, samples of 10 Arabic words in Arabic sign language, their Arabic concepts, their corresponding's English and French signs collected. Images are real and were picked using a mobile camera. Dataset images were not captured in a controlled setting; it was especially prone to differences in light, skin color, and other differences in the environment. A snapshot of a few images from these samples is given in Fig. 7.

Fig. 7. Snapshot of the Dataset.

The proposed system applied also to the pre-made ArSL2018 dataset in [20]. The Arabic Sign Language dataset (ArSL2018) is compiled of 54,049 images for the 32 Arabic sign language sign and alphabets collected from 40 participants in different age groups. The dataset contains grayscale 64 * 64 dimensions.

About 80% of the dataset used in training and 20% used in testing the system.

B. Pre-Process Phase

The pre-processing phase consists of many main operations, which applied to the Arabic singe language images after it is taken from the camera. That is to enhance the image. These operations applied to the training set, testing set and on any new images to be recognized. The pre-process operations are as the following:

- Cropping: cropped out the hand object and discarded all the other unnecessary elements in the image.
- Resizing: The images were in different dimensions; therefore, each image resized to a common resolution of 64 × 64 pixels.
- Converting: All the images were converted into grayscale before feeding into the model.
- Also, the collected dataset images smoothed by convolving with a Gaussian filter and median blur to reduce noise in the images.

C. CNN Training

The input consists of standard RGB images of similar size 64 x 64 pixels. Then the dataset was augmented where the images are randomly rotated 0 to 30 degrees. Images were randomly sheared in a range of 0.2-degrees, width and height shift in a range of 0.1 degrees, zoomed in a range of 0.2 degrees and images were horizontally flipped. 200 sign images of Arabic gestures were used in the training phase. The classification accuracy of the collected training set was 98.6 %.

Also, to evaluate CNN model performances, (42960 images) about 80% of (ArSL2018) sign images of Arabic gestures were used to measure the training rate of the CNN architecture. The classification accuracy of this training set was 98.06%.

D. Testing Phase

The input of the system is the image of the sign language static gesture that needs to be recognized. The saved network parameter is loaded to recognize this sign and produced its corresponding label. Then the ontology layer used to produce all available meanings of this input sign. The method used for all test cases and tabulated the obtained accuracies in Section V.

Also, the input could be a text (Arabic, English or French word) which users type to select all respective signs using ontology-based.

E. Connection Between DL and MSLO Ontology for Semantic Translation

After loading CNN and recognize image gesture the output used as input of the semantic layer. SPARQL is used to predict semantically the corresponding meaning and the corresponding sign gesture in another language. SPARQL is used to access the RDFs inside the ontology by using a SELECT statement as a query. SPARQL query has a standard syntax and depends on using variables that contain the predicate, subject, and objects for RDF.

Also, the input of the system could be a text (Arabic, English or French word) which users type to select their respective signs. The process of retrieving and displaying relevant images is based on the user's queries. SPARQL query used in searching ontology to get meaning that represents this word and gets the URL sign image of this meaning.

Some of SPARQL queries used in searching and retrieving data from MSLO:

- SPARQL query to translate Arabic meaning to its corresponding Arabic sign language.

```
SELECT ?arb ?o
WHERE {?arb rdfs:label ?x.
?arb sign:arabic_sign ?o
```

FILTER (regex(str(?x) , """+arabictext+"""))
bind(strafter(str(?arb),str(sign:)) as ?arb) }

- SPARQL query to translate Arabic sign to English sign.

SELECT ?o ?s ?z
WHERE {?o sign:label ?x.
?o sign:english_meaning ?s.
?s sign:english_sign ?z
FILTER (regex(str(?o) ,"""+label+"""))
bind(strafter(str(?s),str(sign:)) as ?s) }

V. Analysis of Results

The proposed DL model was tested on images that had not seen before. Data (11089 images) about 20% of the pre-made dataset used in testing. The recognition rate is defined by the ratio of the numbers of correct sign recognition to the total number of testing signs. Recognition accuracy was 88.87%. Table I shows the performance of trained network on testing data.

Also, Table I shows Arabic letter name, letter name in English script, the number of trained and tested samples, correctly recognized, and misrecognized samples for each Arabic letter.

- The total number of training samples is 42960.
- The total number of testing samples is 11089.
- The total number of correct recognized signs is 9855 about 88.87% of testing samples.
- The total number of incorrectly recognized signs is 1234 about 11.13% of testing samples.

Also, system performance is evaluated by the rate of successful semantic sign language recognition of new signs such that all possible meanings are listed.

The semantic translation rate in percentages in the case of collected 10 Arabic signs is showed in Table II. Without using Ontology, it might be difficult to find meaningful connections among synonyms signs, words in one language and words in different languages.

Table II shows the number of training and testing samples for each word, incorrectly recognized samples of each sign gestures and results with traditional deep learning model and when using SDL.

Also, Table II shows the number of correctly recognized and misrecognized samples for each Arabic word and the corresponding accuracy. The results also show that while most of the words signs are recognized with 100% accuracy, it can be seen that the system recognition accuracy becomes less when gives the input signs for the words" شكرا"and "سيء".

Overall, out of the 88 Arabic sign images used for testing 83 images were recognized to correct meanings and the remaining 5 images were misrecognized resulting. The recognition accuracy was checked as per the correctness of all testing gestures made was 94.31 %.

The results obtained when recognize sign gesture using deep learning model were only the label that represents the word. But when recognize using deep learning and Ontology, the results were all available meanings of this sign gesture.

TABLE. I. Training and Testing Premade Dataset Samples

Letter Name in English Script	Letter Name in Arabic Script	Training Samples	Testing Samples	Correct Recognized	Incorrect Recognized	Letter Name in English Script	Letter Name in Arabic Script	Training Samples	Testing Samples	Correct Recognized	Incorrect Recognized
Alif	أ (ألف)	1035	637	631	6	Za	ظ (ظاء)	1378	345	345	0
B a	ب (باء)	1432	359	259	100	Ain	ع (عين)	1691	423	357	66
Ta	ت (تاء)	1471	367	332	35	Ghin	غ (غين)	1582	395	389	6
Thaa	ث (ثاء)	1413	353	294	59	Fa	ف (فاء)	1564	391	261	130
Jeem	ج (جيم)	1242	310	295	15	Qaf	ق (قاف)	1364	341	223	118
Haa	ح (حاء)	1211	315	311	4	Kaaf	ك (كاف)	1419	355	232	123
Khaa	خ (خاء)	1268	339	290	49	Laam	ل (لام)	1464	368	340	28
Dal	د (دال)	1308	326	319	7	Meem	م (ميم)	1412	353	325	28
Zal	ذ (ذال)	1265	317	313	4	Nun	ن (نون)	1506	313	213	100
Ra	ر (راء)	1328	331	317	14	Ha	ه (هاء)	1274	318	244	74
Zay	ز (زاي)	1100	274	269	5	Waw	و (الواو)	1096	275	275	0
Seen	س (سين)	1311	327	325	2	Ya	ئ	1377	345	344	1
Sheen	ش (شين)	1205	302	300	2	Taa	ة (تاة)	1433	358	303	55
Sad	ص (صاد)	1516	379	324	55	Al	ال (ال)	1075	268	268	0
Dhad	ض (ضاد)	1336	334	236	98	Laa	لا (لا)	1397	349	299	50
T a	ط (طاء)	1452	364	364	0	Yaa	ي(ياء)	1035	258	258	0
		20893	5634	5179	455			22067	5455	4676	779

TABLE. II. NUMBER OF SAMPLES, RECOGNITION RATE (%) AND RESULTS PER WORD

Arabic Words	Label	Training Samples	Testing Samples	Incorrect Recognized	Result with Traditional DL	Results of Proposed Method	
						Synonyms	*Words in Same sign*
سئ	Bad	16	5	1	Bad	الردِيء, 'السُّوء, السّيِّء, الطالِح	
سيارة	Car	20	8	0	Car	سيّارة, سيْارة, عربة,حافلة, مرْكبة	
جيد	Good	30	22	0	Good	مُتْقَن ,جيد ,حسن	حرف الألف
مرحبا	Hello	19	7	0	Hello	مَرْحَباً , مرحبا	الرقم خمسة
أحبك	I_love_you	17	9	0	I_love_you	اقبلك ,احبك	
الحب	Love	22	7	0	Love	حُبّ, وَجْد, قبول	الاعتذار
لا	No	20	9	0	No	لا ,نفي ,اعتراض	حرف الباء ,الرقم واحد
صورة	Photo	29	6	0	Photo	تِمْثال , خَيَال , شَبَه, شَبِيه , صُورَة , نموزج , مَثِيل , مَجَاز , مِثَال , نَظِير ,صورة فوتوغرافية , تَصْوِيرَة	
شكرا	Thanks	11	8	4	Thanks	اِمْتِنان, عِرْفانُ بالإحْسان, الثَّناءُ الجميلُ , شكرا	
نعم	Yes	16	7	0	Yes	نعم ,موافقة ,قَبُول ,أجل , بلى	
		200	88	5			

VI. CONCLUSION AND FUTURE WORKS

In this paper, the proposed Ontology with Deep Learning (SDL) method bridges the gap between Ontology and deep learning thus taking advantages of both. The proposed Multi-Sign Language Ontology (MSLO) first version to solve some sign language challenges. It extends the WordNet properties to deal with sign language. Some relations were added in MSLO to be compatible with multiple sign languages.

A Pre-made Arabic Sign Language dataset is used to evaluate the built CNN architecture. The training classification accuracy was 98.06 and the recognition accuracy was 88.87%.

To check the feasibility of the proposed method, it was applied to a semantic translation system for translating Arabic static signs to its meanings as Arabic text and other sign languages. Where CNN trained model used in the recognition process with adding the semantic layer. Collected signs of 10 Arabic gestures and their meanings in English and French sign languages used in training and testing the system. Best results were obtained; semantic translation accuracy was 94.31%. Also, the results compared when using deep learning only and when using ontology with deep learning. The SDL output presents all available meanings of the input sign gesture.

As future work, the SDL will enhance to cover dynamic real video in real-time applications. The system will convert to an Android to produce a free long-life mobile application for deaf and speaking impaired people.

REFERENCES

[1] M.M. Reda, N.G. Mohammed ,and R.A.A.A.A. Seoud. SVBiComm: Sign-Voice Bidirectional Communication System for Normal,“Deaf/Dumb” and Blind People based on Machine Learning. 1st International Conference on Computer Applications & Information Security (ICCAIS) ,IEEE, pages 1-8, 2018.

[2] S. Saqib and S. A. R. Kazmi. Repository of static and dynamic signs. International Journal of Advanced Computer Science and Applications(IJACSA), 8(11),2017.

[3] N. B. Ibrahim, M. M. Selim and H. H. Zayed. An automatic arabic sign language recognition system (arslrs). Journal of King Saud University-Computer and Information Sciences, 30(4): 470-477,2018.

[4] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. Nature 521(7553): 436–444, May 2015.

[5] W. Wang, J. Yang, J. Li. S. Xiao, and D. Zhou. Face recognition based on deep learning. In International Conference on Human Centered Computing, Springer, Cham, pages 812-820, 2014.

[6] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata. Audio-visual speech recognition using deep learning. Applied Intelligence, 42(4): 722-737,2015.

[7] A. A. Abu-Shareha and M. Rajeswari. A review on ontology-based label extraction from image data. Journal of Theoretical & Applied Information Technology, 71(2): 268-280,2015.

[8] N.B. Ibrahim, B. Nada, Mazen ,M.M Selim ,and H. H. Zayed. An automatic arabic sign language recognition system (ArSLRS). Journal of King Saud University-Computer and Information Sciences, 30(4) : 470-477,2018.

[9] R. Alzohairi, R.Alghonaim, W.Alshehri, S.Aloqeely, M.Alzaidan and O.Bchir. Image based arabic sign language recognition system. International Journal of Advanced Computer Science and Applications (IJACSA), 9(3), 2018.

[10] M. Abdo, A. Hamdy, S. Salem, and E. M. Saad. Arabic alphabet and numbers sign language recognition. International Journal of Advanced Computer Science and Applications (IJACSA), 6(11):209-214,2015.

[11] S. Hayani, M. Benaddy, O. El Meslouhi, and M. Kardouchi. Arab Sign language Recognition with Convolutional Neural Networks. In 2019 International Conference of Computer Science and Renewable Energies (ICCSRE), IEEE, pages 1-4,2019.

[12] M. ElBadawy, A. S. Elons, H. A.Shedeed, and M. F Tolba. Arabic sign language recognition with 3D convolutional neural networks. In 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS) ,IEEE, pages 66-71, 2017.

[13] A. M. Almasoud, and H. S. Al-Khalifa. SemSignWriting: A Proposed Semantic System for Arabic Text-to-SignWriting Translation. Journal of Software Engineering and Applications, 5(08): 604-612, 2012.

[14] O. Lozynska, and M. Davydov. Information technology for Ukrainian Sign Language translation based on ontologies. ECONTECHMOD: an international quarterly journal on economics of technology and modeling processes , 4(2): 13-18, 2015.

[15] G. Petrucci, C. Ghidini , and M. Rospocher. Ontology learning in the deep. In European Knowledge Acquisition Workshop, Springer, Cham, pages 480-495, 2016.

[16] P. Hohenecker and T. Lukasiewicz. Deep learning for ontology reasoning. arXiv preprint arXiv:1705.10342 ,2017.

[17] J. Zhang and C. Zong. Deep Neural Networks in Machine Translation: An Overview. IEEE Intelligent Systems, 5(30):16-25,2015.

[18] S.Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 ,2015.

[19] L. Perez and J. Wang. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621, 2017.

[20] G. Latif, N. Mohammad, J. Alghaz, R. AlKhalaf and R. AlKhalaf. ArASL: Arabic Alphabets Sign Language Dataset. Data in Brief, Volume 23, Article 103756, 2019.

Malicious URL Detection based on Machine Learning

Cho Do Xuan[1], Hoa Dinh Nguyen[1]
Information Security dept, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam[1, 2]
Information Assurance dept, FPT University, Hanoi, Vietnam[1]

Tisenko Victor Nikolaevich[3]
Systems of automatic Design
Peter the Great St. Petersburg Polytechnic University
Russia, St.Petersburg
Polytechnicheskaya, 29

***Abstract*—Currently, the risk of network information insecurity is increasing rapidly in number and level of danger. The methods mostly used by hackers today is to attack end-to-end technology and exploit human vulnerabilities. These techniques include social engineering, phishing, pharming, etc. One of the steps in conducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a results, malicious URL detection is of great interest nowadays. There have been several scientific studies showing a number of methods to detect malicious URLs based on machine learning and deep learning techniques. In this paper, we propose a malicious URL detection method using machine learning techniques based on our proposed URL behaviors and attributes. Moreover, bigdata technology is also exploited to improve the capability of detection malicious URLs based on abnormal behaviors. In short, the proposed detection system consists of a new set of URLs features and behaviors, a machine learning algorithm, and a bigdata technology. The experimental results show that the proposed URL attributes and behavior can help improve the ability to detect malicious URL significantly. This is suggested that the proposed system may be considered as an optimized and friendly used solution for malicious URL detection.**

Keywords*—*URL; malicious URL detection; feature extraction; feature selection; machine learning

I. Introduction

Uniform Resource Locator (URL) is used to refer to resources on the Internet. In [1], Sahoo et al. presented about the characteristics and two basic components of the URL as: protocol identifier, which indicates what protocol to use, and resource name, which specifies the IP address or the domain name where the resource is located. It can be seen that each URL has a specific structure and format. Attackers often try to change one or more components of the URL's structure to deceive users for spreading their malicious URL. Malicious URLs are known as links that adversely affect users. These URLs will redirect users to resources or pages on which attackers can execute codes on users' computers, redirect users to unwanted sites, malicious website, or other phishing site, or malware download. Malicious URLs can also be hidden in download links that are deemed safe and can spread quickly through file and message sharing in shared networks. Some attack techniques that use malicious URLs include [2, 3, 4]: Drive-by Download, Phishing and Social Engineering, and Spam.

According to statistics presented in [5], in 2019, the attacks using spreading malicious URL technique are ranked first among the 10 most common attack techniques. Especially, according to this statistic, the three main URL spreading techniques, which are malicious URLs, botnet URLs, and phishing URLs, increase in number of attacks as well as danger level.

From the statistics of the increase in the number of malicious URL distributions over the consecutive years, it is clear that there is a need to study and apply techniques or methods to detect and prevent these malicious URLs.

Regarding the problem of detecting malicious URLs, there are two main trends at present as malicious URL detection based on signs or sets of rules, and malicious URL detection based on behavior analysis techniques [1, 2]. The method of detecting malicious URLs based on a set of markers or rules can quickly and accurately detect malicious URLs. However, this method is not capable of detecting new malicious URLs that are not in the set of predefined signs or rules. The method of detecting malicious URLs based on behavior analysis techniques adopt machine learning or deep learning algorithms to classify URLs based on their behaviors. In this paper, machine learning algorithms are utilized to classify URLs based on their attributes. The paper also includes a new URL attribute extraction method.

In our research, machine learning algorithms are used to classify URLs based on the features and behaviors of URLs. The features are extracted from static and dynamic behaviors of URLs and are new to the literature. Those newly proposed features are the main contribution of the research. Machine learning algorithms are a part of the whole malicious URL detection system. Two supervised machine learning algorithms are used, Support vector machine (SVM) and Random forest (RF).

The paper is organized as follows. Section II reviews some recent works in the literature on malicious URL detection. The proposed malicious URLs detection system using machine learning is presented in Section III. In this section, the new features for URLs detection process are also described in details. Experimental results and discussions are provided in Section IV. The paper is concluded by Section V.

II. Related Works

A. Signature based Malicious URL Detection

Studies on malicious URL detection using the signature sets had been investigated and applied long time ago [6, 7, 8]. Most of these studies often use lists of known malicious URLs. Whenever a new URL is accessed, a database query is executed. If the URL is blacklisted, it is considered as malicious, and then, a warning will be generated; otherwise URLs will be considered as safe. The main disadvantage of this approach is that it will be very difficult to detect new malicious URLs that are not in the given list.

B. *Machine Learning based Malicious URL Detection*

There are three types of machine learning algorithms that can be applied on malicious URL detection methods, including supervised learning, unsupervised learning, and semi-supervised learning. And the detection methods are based on URL behaviors.

In [1], a number of malicious URL systems based on machine learning algorithms have been investigated. Those machine learing algorithms include SVM, Logistic Regression, Nave Bayes, Decision Trees, Ensembles, Online Learning, ect. In this paper, the two algorithms, RF and SVM, are used. The accuracy of these two algorithms with different parameters setups will be presented in the experimental results.

The behaviors and characteristics of URLs can be divided into two main groups, static and dynamic. In their studies [9, 10, 11] authors presented methods of analyzing and extracting static behavior of URLs, including Lexical, Content, Host, and Popularity-based. The machine learning algorithms used in these studies are Online Learning algorithms and SVM. Malicious URL detection using dynamic actions of URLs is presented in [12, 13]. In this paper, URL attributes are extracted based on both static and dynamic behaviors. Some attribute groups are investigated, including Character and semantic groups; Abnormal group in websites and Host-based group; Correlated group.

C. *Malicious URL Detection Tools*

- URL Void: URL Void is a URL checking program using multiple engines and blacklists of domains. Some examples of URL Void are Google SafeBrowsing, Norton SafeWeb and MyWOT. The advantage of the Void URL tool is its compatibility with many different browsers as well as it can support many other testing services. The main disadvantage of the Void URL tool is that the malicious URL detection process relies heavily on a given set of signatures.
- UnMask Parasites: Unmask Parasites is a URL testing tool by downloading provided links, parsing Hypertext Markup Language (HTML) codes, especially external links, iframes and JavaScript. The advantage of this tool is that it can detect iframe fast and accurately. However, this tool is only useful if the user has suspected something strange happening on their sites.
- Dr.Web Anti-Virus Link Checker: Dr.Web Anti-Virus Link Checker is an add-on for Chrome, Firefox, Opera, and IE to automatically find and scan malicious content on a download link on all social networking links such as Facebook, Vk.com, Google+.
- Comodo Site Inspector: This is a malware and security hole detection tool. This helps users check URLs or enables webmasters to set up daily checks by downloading all the specified sites. and run them in a sandbox browser environment.
- Some other tools: Among aforementioned typical tools, there are some other URL checking tools, such as UnShorten.it, VirusTotal, Norton Safe Web, SiteAdvisor (by McAfee), Sucuri, Browser Defender, Online Link Scan, and Google Safe Browsing Diagnostic.

From the analysis and evaluation of malicious URL detection tools presented above, it is found that the majority of current malicious URL detection tools are signature-based URL detection systems. Therefore, the effectiveness of these tools is limited.

III. MALICIOUS URL DETECTING USING MACHINE LEARNING

A. *The Model*

Fig. 1 presents the proposed malicious URL detection system using machine learning. The malicious URL detection model using machine learning contains two stages: training and detection.

- Training stage: To detect malicious URLs, it is necessary to collect both malicious URLs and clean URLs. Then, all the malicious and clean URLs are correctly labeled and proceeded to attribute extraction. These attributes will be the best basis for determining which URLs are clean and which are malicious. Details of these attributes will be presented in details in this paper. Finally, this dataset is divided into 2 subsets: training data used for training machine learning algorithms, and testing data used for testing process. If the classification performance of the machine learning model is good (high classification accuracy), the model will be used in the detection phase.
- Detection phase: The detection phase is performed on each input URL. First, the URL will go through attribute extraction process. Next, these attributes are input to the classifier to classify whether the URL is clean or malicious.

B. *URL Attribute Extraction and Selection*

In [1], the authors listed some main attribute groups for malicious URL detection as follows.

Lexical features: these features include URL length, main domain length, maximum token domain length, path average length, average token length in domain.

Host-based Features: these features are extracted from the host characteristics of the URLs. These attributes indicate the location of malicious servers, the identity of malicious servers, the degree of impact of several host-based features that contribute the URL's malicious level.

Content-based Features: these features are acquired when a whole web page is downloaded. The workload of these features is quite heavy, since a lot of information needs to be extracted, and there may be security concerns about accessing that URL. However, with more information available about a particular site, it is expected to create a better prediction model. The content-based features of a website can be extracted primarily from its HTML content and the use of JavaScript.

Above are the three main attribute groups commonly used by researchers to detect malicious URLs. However, each study has its own decision on suitable attributes and characteristics for each particular experimental dataset. In this paper, the use of all three attribute groups is recommended. However, in each attribute group some new attributes and characteristics of the URL to optimize the ability to detect malicious URLs are proposed. The new attributes for malicious URL detection in this research are listed in Tables I, II, and III.

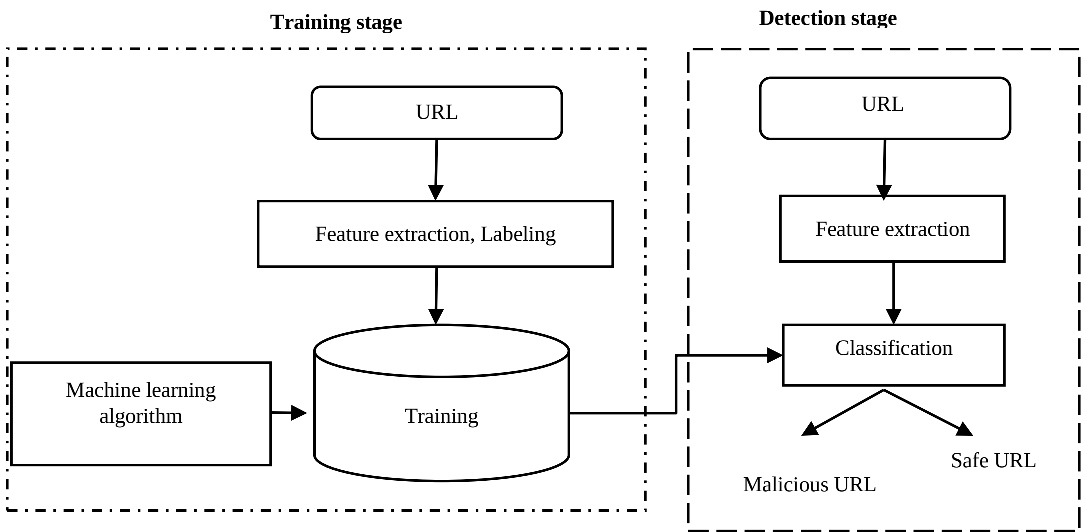


Fig. 1. Malicious URL Detection Model using Machine Learning.

TABLE. I. LIST OF URL FEATURES IN LEXICAL FEATURE GROUP

No	Feature group	Feature	Data type	Description
1	Lexical group	NumDots	numeric	Number of character '.' in URL
2		SubdomainLevel	numeric	Number of subdomain levels
3		PathLevel	numeric	The depth of URL
4		UrlLength	numeric	The length of URL
5		NumDash	numeric	Number of the dash character '-'
6		NumDashInHostname	numeric	Number of dash character in the hostname
7		AtSymbol	boolean	There exists a character '@' in URL
8		TildeSymbol	boolean	There exists a character '~' in URL
9		NumUnderscore	numeric	Number of the underscore character
10		NumPercent	numeric	Number of the character '%'
11		NumQueryComponents	numeric	Number of the query components
12		NumAmpersand	numeric	Number of the character '&'
13		NumHash	numeric	Number of the character '#'
14		NumNumericChars	numeric	Number of the numeric character
15		NoHttps	boolean	Check if there exists a HTTPS in website URL
16		IpAddress	boolean	Check if the IP address is used in the hostname of the website URL
17		DomainInSubdomains	boolean	Check if TLD or ccTLD is used as a part of the subdomain in website URL
18		DomainInPaths	boolean	Check if TLD or ccTLD is used in the link of website URL
19		HttpsInHostname	boolean	Check if HTTPS is disordered in the hostname of website URL
20		HostnameLength	numeric	Length of hostname
21		PathLength	numeric	Length of the link path
22		QueryLength	numeric	Length of the query
23		DoubleSlashInPath	boolean	There exists a slash '//' in the link path
24		NumSensitiveWords	numeric	Number of sensitive words (i.e., “secure”, “account”, “webscr”, “login”, “ebayisapi”, “sign in”, “banking”, “confirm”) in website
25		EmbeddedBrandName	boolean	There exists a brand name in the domain
26		PctExtHyperlinks*	float	The percentage of external hyper links in the HTML source code of website

TABLE. II. LIST OF URL FEATURE IN THE HOST-BASED FEATURE GROUP

No	Feature group	Feature	Data type	Description
27	Host-based feature group	PctExtResourceUrls*	float	Percentage of URL external resource in HTML source codes of website
28		ExtFavicon*	boolean	Check if favicon is installed from a hostname different from the URL hostname of website
29		InsecureForms*	boolean	Check if actions in the form containing the contend of URL without HTTPS protocol
30		RelativeFormAction*	boolean	Check if the action form contains a relative URL
31		ExtFormAction*	boolean	Check if the action form contains an external URL
32		AbnormalFormAction*	boolean	Check if the action form contains an abnormal URL.
33		PctNullSelfRedirectHyperlinks*	float	Percentage of hyperlinks containing an empty value, an auto-redirecting value, such as "#", URL of current website, or some abnormal values such as "file://E:/"
34		FrequentDomainNameMismatch	boolean	Check if the most frequent hostname in the HTML source code does not match the URL of website.
35		FakeLinkInStatusBar*	boolean	Check if HTML source code contains a JavaScript command on MouseOver to display a fake URL in the status bar
36		RightClickDisabled	boolean	Check if HTML source code contains a JavaScript command to turn off the right click of the mouse
37		PopUpWindow	boolean	Check if HTML source code contains a JavaScript command to start a popup window
38		SubmitInfoToEmail	boolean	Check if HTML source code contains "mailto" in the HTML
39		IframeOrFrame	boolean	Check if iframe or frame is used in HTML source codes
40		MissingTitle	boolean	Check if the title tag is empty in HTML source codes
41		src_eval_cnt	int	Number of function eval () in HTML source codes
42		src_escape_cnt	int	Number of function escape () in HTML source codes
43		src_exec_cnt	int	Number of function exec() in HTML source codes
44		src_search_cnt	int	Number of function search() HTML source codes
45		ImagesOnlyInForm*	boolean	Check if actions in the form of HTML source code does not contain text, but only images
46		rank_country	Boolean	Current country rank of website URL is in top 1 million of Alexa
47		rank_host	Boolean	The rank of the host website URL is in top 1 million of Alexa
48		AgeDomain	int	The age of domain since it is registered

TABLE. III. LIST OF URL FEATURES IN CORRELATED FEATURE GROUP

No	Feature group	Feature	Data type	Description
49	correlated feature group	UrlLengthRT*	-1, 0, 1	Correlated length of URL
50		PctExtResourceUrlsRT*	-1, 0, 1	Correlated percentage of external URL
51		AbnormalExtFormActionR*	-1, 0, 1	Correlated abnormal actions in form
52		ExtMetaScriptLinkRT*	-1, 0, 1	Correlated meta script link
53		SubdomainLevelRT*	-1, 0, 1	Correlated sub-domain level
54		PctExtNullSelfRedirectHyperlinksRT *	-1, 0, 1	Correlated null self-redirect hyperlinks

All attributes marked "*" in Tables I, II, III are newly extracted and selected in this research. Besides, in previous researches, authors tend to use feature extraction and selection method based on a group of predefined features. However, those recommended features are specialized and not popular. As a results, it is usually difficult to implement those features in other works, and to re-evaluate the detection performance of those features. In this work, we try to combine basic features to formulate new ones.

C. Machine Learning Algorithm Selection

The application of machine learning algorithms in detecting malicious URLs has been studied and applied widely [1]. In this paper, two commonly used supervised machine learning algorithms, RF and SVM [14, 15], are used.

In this research, machine learning algorithms are the last puzzle to complete our proposed malicious URL detection system. Those algorithms are suitable to utilized the usefulness of our new features selected for malicious URL detection. The machine learning algorithms are already well investigated in the literature. In this work, SVM and RF are selected as an example to illustrate the good performance of the whole detection system, and are not our main focus. Readers are encouraged to implement some other algorithms such as Naïve Bayes, Decision trees, k-nearest neighbors, neural networks, etc.

In order to explore the effectiveness of using these two algorithms, different adjustments of parameters are implemented.

IV. EXPERIMENTAL RESULTS

A. Dataset and Experiment Environments

1) Experiment dataset: The experimental dataset for malicious URL detection model includes: 470.000 URLs collected from [16, 17, 18, 19], of which about 70.000 URLs are malicious and 400.000 URLs are safe. All these URLs are checked by Virus Total tool to verify the labels of each URL. The complete dataset is stored using CSV format. Each URL sample has a label "bad" for malicious and "good" for safe. Details of the data are as follows:

- Phishtank [16]: Phishtank is a service Website dedicated for sharing phishing URLs. Suspicious URLs can be sent to Phishtank for verification. The data in Phishtank is updated hourly.
- URLhaus [17]: URLhaus is a project from abuse.ch aiming at sharing malicious URLs being used for malicious software distribution.
- Alexa [18]: Is a database ranking all websites according to their usefulness.
- Malicious_n_Non-Malicious URL [19]: is a data source with more than 400,000 labeled URL. In this database, 82% of all URLs are safe, while remaining 18% of URLs are malicious.

2) Experimental setup: The dataset of both safe and malicious URLs mentioned above is divided into 2 subsets. About 80% of the dataset, 470.000 URLs (400.000 safe URLs, 70.000 malicious URL), is used for training, and about 20% of the dataset, about 10.000 URLs (5.000 malicious URLs, 5.000 safe URLs), is used for testing. The experiment is repeated many times with both SVM and RF algorithm. Different parameter settings are used in different runs.

3) Experiment dataset

- Setup environment: Python version 3.6; Spark version 2.3.0; Hadoop version 2.7; Java (JDK) 8; Ubuntu 18.04.
- Hardware: RAM 16GB; Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz.

B. Results and Discussions

1) Evaluation metrics: Accuracy: the percentage of correct decisions among all testing samples

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{1}$$

where: *TP*- True positive is the number of malicious URLs correctly labeled; *FN* - False negative is the number of malicious URLs misclassified as safe; *TN*- True negative is the number of safe URL correctly labeled; *FP* - False positive is the number of safe URLs misclassified as malicious.

Confusion matrix: is a two-way Table IV representing how many samples are classified into which label accordingly.

Precision: is the percentage of malicious URLs correctly labeled (TP) among all malicious URLs labeled by the classifier (TP+FP).

$$precision = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

Recall: is the percentage of malicious URLs correctly labeled (*TP*) among all malicious URLs of the testing data (*TP*+*FN*).

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

F1-score: is the harmonic mean of precision and recall. High F1 value means the classifier is good.

$$F1 = \frac{2 \times precision \times \mathrm{Re}\,call}{precision + \mathrm{Re}\,call} \tag{4}$$

FPR (False prediction rate) is calculated as:

$$FRP = \frac{FP}{FP + TN} \times 100\% \tag{5}$$

2) Results

- Training performance

To evaluate the training performance of the machine learning algorithm, both two data subsets are used individually. Each of these data subsets has different data size as well as different distribution of data labels, which may result in different training performances. The results are presented in Table V.

Experimental results show that the RF with 100 trees gives the best predictive result. In return, the training time of the RF is slightly longer than SVM, but the testing time is not much different. The accuracy of the second dataset is reduced due to the unbalance between safe and malicious URLs of the data. As expected, RF algorithm, with its fast speed and high accuracy, is very suitable for classification problem. Besides, in our research, when machine learning algorithms are combined with spark libraries, the training and testing time can be reduced significantly. SparkML Machine Learning is a library package that provides and supports many machine learning algorithms such as SVM, RF, Naïve Bayes, Regression, Clustering, Collaborative Filtering, ... It is a suitable tool for applying machine learning algorithms with fast and accurate processing speed on large datasets.

- Testing results: In this paper, additional small testing dataset, with 107 safe URLs and 118 malicious URLs, is used to evaluate the performance of the best machine learning algorithm discussed above, RF (100). The results are presented in Table VI.

Confusion matrix parameters: TP: 92.174%; FPR: 12.037%; TN: 87.963%; FN: 7.826%

TABLE. IV. CONFUSION MATRIX

	Classified malicious URL	**Classified safe URL**
Real malicious URLs	TP	FN
Real safe URLs	FP	TN

TABLE. V. TRAINING PERFORMANCE OF MALICIOUS URL DETECTION SYSTEM

Dataset	Algorithm and parameters	Accuracy (%)	Precision (%)	Recall (%)	Training time (s)	Testing time (s)
10.000 URLs	SVM (100 iterations)	93.39	94.67	92.51	2.32	0.01
	SVM (10 iterations)	93.35	94.84	92.71	3.11	0.01
	RF (10 trees)	99.10	98.43	97.45	2.78	0.01
	RF (100 trees)	99.77	98.75	97.85	3.34	0.01
470.000 URLs	SVM (100 iterations)	90.70	93.43	88.45	272.97	2.12
	SVM (10 iterations)	91.07	93.75	88.85	280.33	2.31
	RF (10 trees)	95.45	90.21	95.12	372.97	2.02
	RF (100 trees)	96.28	91.44	94.42	480.33	2.30

TABLE. VI. TESTING RESULTS

	Predicted safe URL	Predicted malicious URL
Real safe URL (107)	96	11
Real malicious URL (118)	9	109

V. CONCLUSIONS

In this paper, a method for malicious URL detection using machine learning is presented. The empirical results in Tables V and VI have shown the effectiveness of the proposed extracted attributes. In this study, we do not use special attributes, nor do we seek to create huge datasets to improve the accuracy of the system as many other traditional publications. Here, the combination between easy-to-calculate attributes and big data processing technologies to ensure the balance of the two factors is the processing time and accuracy of the system. The results of this research can be applied and implemented in information security technologies in information security systems. The results of this article have been used to build a free tool [20] to detect malicious URLs on web browsers.

REFERENCES

[1] D. Sahoo, C. Liu, S.C.H. Hoi, "Malicious URL Detection using Machine Learning: A Survey". CoRR, abs/1701.07179, 2017.

[2] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: a literature survey," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.

[3] M. Cova, C. Kruegel, and G. Vigna, "Detection and analysis of driveby-download attacks and malicious javascript code," in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 281–290.

[4] R. Heartfield and G. Loukas, "A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks," ACM Computing Surveys (CSUR), vol. 48, no. 3, p. 37, 2015.

[5] Internet Security Threat Report (ISTR) 2019–Symantec. https://www.symantec.com/content/dam/symantec/docs/reports/istr-24-2019-en.pdf [Last accessed 10/2019].

[6] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in Proceedings of Sixth Conference on Email and Anti-Spam (CEAS), 2009.

[7] C. Seifert, I. Welch, and P. Komisarczuk, "Identification of malicious web pages with static heuristics," in Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian. IEEE, 2008, pp. 91–96.

[8] S. Sinha, M. Bailey, and F. Jahanian, "Shades of grey: On the effectiveness of reputation-based "blacklists"," in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on. IEEE, 2008, pp. 57–64.

[9] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying suspicious urls: an application of large-scale online learning," in Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009, pp. 681–688.

[10] B. Eshete, A. Villafiorita, and K. Weldemariam, "Binspect: Holistic analysis and detection of malicious web pages," in Security and Privacy in Communication Networks. Springer, 2013, pp. 149–166.

[11] S. Purkait, "Phishing counter measures and their effectiveness– literature review," Information Management & Computer Security, vol. 20, no. 5, pp. 382–420, 2012.

[12] Y. Tao, "Suspicious url and device detection by log mining," Ph.D. dissertation, Applied Sciences: School of Computing Science, 2014.

[13] G. Canfora, E. Medvet, F. Mercaldo, and C. A. Visaggio, "Detection of malicious web pages using system calls sequences," in Availability, Reliability, and Security in Information Systems. Springer, 2014, pp. 226–238.

[14] Leo Breiman.: Random Forests. Machine Learning 45 (1), pp. 5- 32, (2001).

[15] Thomas G. Dietterich. Ensemble Methods in Machine Learning. International Workshop on Multiple Classifier Systems, pp 1-15, Cagliari, Italy, 2000.

[16] Developer Information. https://www.phishtank.com/developer_info.php. [Last accessed 11/2019].

[17] URLhaus Database Dump. https://urlhaus.abuse.ch/downloads/csv/. [Ngày truy nhập 11/2019].

[18] Dataset URL. http://downloads.majestic.com/majestic_million.csv. [Last accessed 10/2019].

[19] Malicious_n_Non-MaliciousURL. https://www.kaggle.com/antonyj453/urldataset#data.csv. [Last accessed 11/2019].

[20] chrome.zip. https://drive.google.com/file/d/13G_Ndr4hMFx_qWyTEjHuOyJmHFWD0Gud/view?fbclid=IwAR0SLVCrvjHHGmoHZH97nXN3Bm-DMY7jG4SOsKZYLAZjTFgeoJADfli64-g. [Last accessed 12/2019].

Noise Reduction in Spatial Data using Machine Learning Methods for Road Condition Data

Dara Anitha Kumari[1]
CSE Department, JNTUH CEH
Hyderabad, Telangana, India

Dr. A. Govardhan[2]
CSE Department, JNTUH
Hyderabad, Telangana, India

***Abstract*—With the increase in the road transportation system the safety concerns for the road travels are also increasing. In order to ensure the road safety, various government and non-government efforts are visible to maintain the road quality and transport network system. The maintenance of the road condition is in the verse of getting automated for the quick identification of potholes, cracks and patch works and repair. The automation process is taking place in majority of the counties with the help of ICT enabled frameworks and devices. The primary device used for the purpose is the geo location enabled image capture devices. Regardless to mention the image capture process is always prune to noises and must be removed for better further analysis. Also, the spatial data is collected from the road networks are also prune to various error such as missing values or outliers due to the induced noises in the capture devices. Hence, the demand of the current research is to purpose a complete solution for the noise identification and removal from the spatial road network data for making the automation process highly successful and highly accurate. In the recent time, many parallel research attempts are observed, which resulted into solving the problem of noise reduction in all aspects of spatial data. Nevertheless, all the parallel research outcomes have failed to provide a single solution for all the noise issues. Henceforth, this work proposes three novel algorithms to solve spatial image noise problem using the adaptive moment filtration, missing value noise from the spatial data using adaptive logistic analysis and finally, the outlier noise removal from the same spatial data using corrective logistic machine learning method. The outcome of this work is nearly 70% accuracy in image noise reduction, 90% accuracy for missing value and outlier removal. The work also justifies the information loss reduction by nearly 50%. The final outcome of the work is to ensure higher accuracy for road maintenance automation**.

***Keywords*—Spatial image moments; adaptive logistic de-noising; machine learning; noise removal; correlative corrections**

I. Introduction

Spatial data retrieval for the imaging methods are highly beneficial for detection and provide automatic maintenance of the road conditions as the spatial data provides higher order information for each pixel in the spatial image. The work by D. A. Landgrebe [1] have significantly proven the fact the improvement in further processing can be achieved using spatial data. Nonetheless, the accumulation of the spatial image data is full of challenges as suggested by J. M. Bioucas – Dias [2] with the restriction of continuous mapping between image data and spatial data vectors. The other parallel research by N. Keshava [3] have suggested various other methods for mapping without the spatial mixing methods.

However, there are multiple parallel research outcomes, which suggested extraction of the spatial data by separating the image information and the text information associated with each pixel. The study by M. E. Winter [4] Have suggest the extraction of the road spatial data using the N-Finder algorithm. This algorithm is sophisticated and highly accepted by the researcher for spatial data extraction using the vector methods. Also, the work by J. M. P. Nascimento [5] have demonstrated the vertex component analysis method, which relies on the modifications and enhancements over the principle component extraction methods. Further, the work by J. Li [6] has showcased the algorithm as minimum volume measure, ensuring the extraction of the minimal spatial data for any purpose such as road information.

Nonetheless, having the capture process sorted out for the spatial data accumulation, the major focus is on the noise reduction as suggested by Alp Erturk [7]. Henceforth, this work proposes a noval solution for spatial data noise reduction for all possible noise types.

The rest of the work is furnished such that in Section 2, the parallel research outcomes are analysed, in Section 3, the mathematical model for the identified problems and the proposed solutions are furnished, in Section 4 the proposed algorithm are elaborated, in Section 5, the obtained results are discussed, in Section 6 the comparative analysis is summarized and the work produces the final conclusion in Section 7.

II. Parallel Research Outcomes

In the section of the work, the parallel research outcomes are discussed critically for better formulation of the problem and solutions in the upcoming sections.

The noise in any form in any data can be catastrophic for achieving the further results. Hence, the pre-processing techniques for the data noise reductions are getting popular. However, the present pre-processing method for the spatial data is not highly complex and demands further improvements. When there is significantly immaterial and redundant and data present or noisy and unreliable data, then comprehension discovery during the training period is more difficult steps that are filtering and data preparation may require considerable level of processing period. Data pre-processing comprises cleansing, Instance selection, normalization, transformation, characteristic extraction along with variety, etc.

The recent research by M.Zortea [8] have pro-posed a spatial pre-processing method for image data using endmember extraction process. The work is highly justified for the pixel information preservation and noise reduction at the same time. Nonetheless, the associated information extracted from the pixels are also prune to noises, which must be de-noised as well. However, the work by G.Martin [9] de-noising process, where the relevancy of the information is also considered. This ensures some of the extracted information from te pixels are also processes for normalization. Further, the work of A. Plaza [10] has fine-tuned the workability of the spatial image information de-noising by applying the remotely sensed hyperspectral analysis for the spatial datasets.

In the contrast, the work by A. Erturk [11] and D. Cesmeci [12] has demonstrated significant contributions to the de-noising of the spatial data. These two works are highly accepted. Nevertheless, this work identified few drawbacks in these two bench-marked algorithms and proposes a solution to these methods. In the further sections of the work, these two works are constantly compared with the proposed outcomes. Finally, the similar problems are also highlighted by the work of C. Li [13].

The limitations of the existing research outcomes are listed here: Firstly, all signal processing apparatus, both analogy along with electronic, have faculties which make them more prone to noise. Noise might be arbitrary or white sound using an additional frequency supply, or frequency determined noise introduced with means of a computer device's mechanism or indicate processing calculations.

Secondly, back in electronic recording apparatus, a significant kind of noise can be hiss created by arbitrary electron motion due to thermal agitation in any way temperatures above absolute zero. These abrasive electrons quickly add and subtract out of the voltage of this output and therefore create detectable sound.

Finally, in the instance of photographic picture along with magnetic tape is introduced on account of the grain structure of this moderate. In photographic picture, the magnitude of the sausage at the film determines the picture's sensitivity, more sensitive picture with larger sized proteins. To pay for this, larger regions of film or magnetic tape could be employed to lessen the noise to a decent level.

Henceforth, in the next section of the work, the problems are highlighted, elaborated and the mathematical models for the proposed solutions are formulated.

III. Problem Formulation

After the detailed analysis of the parallel research outcomes, in the section of the work, the mathematical model for the problem and proposed solutions are elaborated.

Lemma-1: Adaptive Moment Filtration can detect and reduce the noise in the spatial image sets.

Proof: Assuming that the complete spatial image ste can be represented as I[] and each and every image in the set can be identified as I_x with a total number of n images in the dataset. Then, this analogy can be represented as:

$$I[] \leftarrow \sum_{x=1}^{n} I_x \tag{1}$$

Where each I_x belongs to the spatial dataset. As,

$$I_x \in I[] \tag{2}$$

Also, every image in the dataset can be represented using the<r,g,b> tuple as,

$$I_x = < r_x, g_x, b_x > \tag{3}$$

Applying, the ϕ function for extraction of the gray parameter as intensity of the image as g_x

$$\phi(I_x) \rightarrow g_x \tag{4}$$

Henceforth the complete image can be clustered based on the adaptive image intensity. Assuming the cluster collection for the image as C [], the following formulation can be derived:

$$C[] = \sum_{i,j=1}^{n.m} \frac{(g_i - g_{i+1})^n.(g_j + g_{j-1})^m}{\{(g_{j-1} - g_j) + (g_{j+1} - g_j) + (g_{i-1} - g_i) + (g_{i+1} - g_i)\}} \tag{5}$$

Further, assuming that, each and every cluster in the cluster set C[] can be represented in terms of two dimensional function to map into the spatial space, then this concept can be mathematical presented as,

$$C_k \in C[] \tag{6}$$

And,

$$C_k = \{ f(i,j) \,|\, (m,n) \} \tag{7}$$

Where m and n are the order of the spatial data. Thus the moment M_k, can be calculated as,

$$M_k = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} i^m j^n . f(i,j).di.dj \tag{8}$$

Similarity, for the complete image, the same moment function can be calculated as

$$M_I = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} i'^{m'} j'^{n'} . f(i',j').di'.dj' \tag{9}$$

Where,

$$C_I = \{ f(i',j') \,|\, (m',n') \} \tag{10}$$

Finally, building the complete set of clusters with noises CN [], each and every cluster must be analysis as,

$$CN[] = \forall C_k \in C[], iff\ M_k \neq M_I \tag{11}$$

Assuming that each cluster in the image set with noise can be represented as CN_n , then each CN_n must be clustered again using the adaptive clustering process and the moments shall be calculated, considering only the selected cluster.

Considering the final size of the sub-cluster is relatively small considering the complete image, the sub-cluster with the noise can be replaced using the adaptive missing value replacement method. This adaptive missing value replacement method is further elaborated in the text mathematical model in this section.

Lemma -2: Adaptive logistic analysis for the missing value can be highly accurate compared with the linear analysis for spatial data.

Proof: Assuming that, the spatial dataset can be considered as D [] and each item in the data set can be represented as D_x with total number of elements as n with total m number of tuples. Thus, this relation can be formulated as,

$$D[] = \overline{\sum_{x=1}^{m} < D_x, D'_x, D''_x ... >}^{n} \tag{12}$$

The linear method applies a simple strategy for calculating the replacement factor, RF [] as,

$$RF[] = \sum_{x=1}^{n^m} | D_{x+1} - D_x | \tag{13}$$

Henceforth, the men of the RF [], denoted as rf, can be calculated as,

$$rf = \frac{\sum_{i=1}^{n^m} RF[i]}{\delta(RF[])} \tag{14}$$

In the other hand, the adaptive logistic analysis builds the correlation factor, Corr, before calculating the rectification factor as,

$$Corr(t+1) = \sum_{x=1}^{n^m} \frac{\partial^2 \{\ln(D_{x+1}) - \dfrac{\ln(D_x)}{\Delta Corr(t)}\}}{\partial \dfrac{\ln(D_{x+1})}{\Delta Corr(t)}^2} \tag{15}$$

Further, the rectification factor, rf, can be calculated as,

$$rf' = \frac{\sum_{t=1}^{n^m} Corr(t)}{\delta(Corr[])} \tag{16}$$

It is natural to realize that, the adaptive logistic method is prone to be a lesser number compared to the linear method [Fig. 1], Hence, the following statement can be made, as

$$rf' << rf \tag{17}$$

Henceforth, if any data point can be featured as missing value data point in the spatial dataset,

$$D_x \to (N / A) \to 0 \tag{18}$$

And

$$D_x \notin D_{x-1} \pm Corr[] \tag{19}$$

Then, using the adaptive logistic method, the missing data point can be calculated as,

$$D_x(t+1) = D_{x-1}(t) \pm Corr[t] \tag{20}$$

Thus, it is realist to the state that, adaptive logistic method is more accurate for identifying and replacing the missing values or data points in the spatial datasets.

Furthermore, the final identified challenge for the spatial data noise reduction is the outlier removal from the outlier form dataset or data points. Hence, this problem must be addressed and the solution to this problem is formulated using the following mathematical model.

Lemma -3 Corrective logistic analysis for the outlier value can be accurate compared with the linear analysis for spatial data.

Proof: Assuming that, the spatial dataset can be considered as D [] and each time in the data set can be represented as D_x with total number of elements as n with total m number of tuples. Thus, this relation can be formulated as,

$$D[] = \overline{\sum_{x=1}^{m} < D_x, D'_x, D''_x ... >}^{n} \tag{21}$$

The liner method applies a simple strategy for calculating the replacement factor, RF [] as,

$$RF[] = \sum_{x=1}^{n^m} | D_{x+1} - D_x | \tag{22}$$

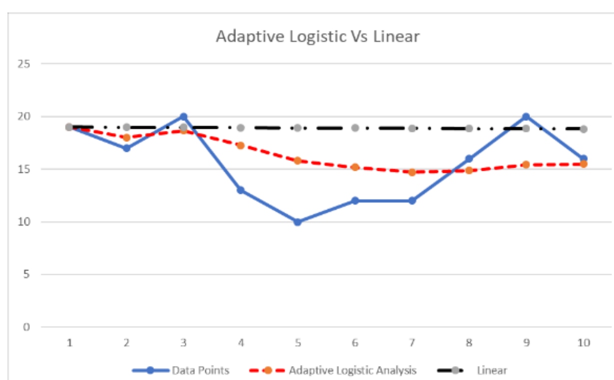


Fig. 1. Comparative Function Growth in Adaptive Logistic and Linear.

Henceforth, the mean of the RF [], denoted as rf, can be calculated as

$$rf = \frac{\sum_{i=1}^{n^m} RF[i]}{\delta(RF[])} \tag{23}$$

In the other hand, the adaptive logistic analysis builds the threshold factor, TH, as,

$$TH(t) = \frac{\sum_{i=0}^{n} \ln(D_x)^x}{\partial t} \pm \theta(t) \tag{24}$$

Further, the correction factor, $\theta(t)$, can be calculated as,

$$\theta(t) = \frac{\theta(t-1)}{\partial\theta[]} \tag{25}$$

Again, it is natural to realize that corrective logistic method is prone to be a lesser number compared to the linear method and can be formulated as

$$TH << rf \tag{26}$$

Henceforth, if any data point can be featured as outlier data points in the spatial dataset,

$$D_x(t) \rightarrow TH(t+1) \tag{27}$$

Then, the outlier data point must be replaced with the newer data point as,

$$D_x(t) \rightarrow TH(t) \pm \theta(t) \tag{28}$$

Thus, it is realist to state that, corrective logistic method is more accurate for identifying and replacing the outliers' values or data points in the spatial datasets.

Henceforth, in the light of the problem formulation and proposed corrective models, in the next section of this work, the proposed algorithms are furnished and discussed.

IV. Proposed Algorithms

After the detailed mathematical modelling of the problem and proposed solution, in this section of the work, the proposed algorithms are furnished and discussed.

Firstly, the algorithm for noise detection and reduction is furnished here for the spatial images.

Moments are well-known for his or her own application in picture investigation, because they are sometimes utilized to derive invariants with regard to special conversion lessons. The expression invariant minutes can be abused within this circumstance. But whilst instant invariants are all invariants which can be made by minutes, the single minutes which can be invariants on their own would be the fundamental minutes. Be aware the invariants will be invariant from the domain names. Neither scaling nor spinning have been characterized also the transformation is not reversible, and also an image is an approximation. All these invariants are simply invariant after describing a silhouette.

Secondly, the algorithm for missing value reduction for spatial data is furnished here.

Algorithm - I: Adaptive Moment Based Spatial Image Noise Detection and Removal Algorithm (**AMBSI-NDR**)

Step - 1. Accept the spatial data set V [].

Step - 2. For each V [i] in V[]

- a. Separate the text component as T[i] and Image Components as I[i].
- b. For each T[i] in T[]
  - i. Call the ALC-MVIR and CLC-OIR algorithms.
- c. For each I[k] in I[]
  - i. Calculate the Pixel Intensity for all Pixels, PI [].
  - ii. Calculate the Initial Noise Level as NI.
  - iii. Calculate the Image Size as IS.
  - iv. For each PI[i] in PI[]
    1. Consider the Image as ImgTem.
    2. Identify the Image Segments as S[].
  - v. Calculate the Image Moment as IM.
  - vi. For each S[i] in S[]
    1. Calculate the moment as M[i]
    2. If M[i] Not Equals to IM
       - a. Then, ImgTem = S[i], IM = M[i] and Repeat from Step-2.C.I
    3. Else, Mark the S[i] as No-Noise
  - vii. Replace the Segments Marked No-Noise data using CLC-OIR Algorithm
- d. Calculate the Reduced Noise Level as RNI.
- e. Calculate the Final Image Size as FIS.
- f. If RNI < NI and FIS = IS
  - i. Then, Report the Final Image without Noise.
- g. Else, Repeat from Step-2.C.

Step - 3. Report the final noise reduced image set I[].

Imputation could be the procedure for substituting lost data using values. It’s called thing imputation when it’s called imputation. When substituting for a factor of the data level. Now, there are three major difficulties that data will cause data make reductions, create the analysis and handling of their data arduous, and may present a sizable total of prejudice.

Algorithm - II: Adaptive Logistic Correlation Based Missing Value Identification & Replacement Algorithm (**ALC-MVIR**)

Step - 1. Accept the Text set as T[] from the Spatial Dataset

Step - 2. For each T[i] from T[] set

a. Calculate the Correlation, Corr, for each Data Iteams as Eq. 15

b. Calculate the rectification factor, Corr as Eq. 16

c. If T[i] Equals to Zero or T[i] Equals to "Not Available"

d. Then,

i. Mark the data items or Data Points as Missing Value

ii. Replace the Missing Value as T[i] = T[i-1] ± Corr[i-1]

Step - 3. Report the final Missing Value Reduced dataset as T[]

Finally, the algorithm for outlier detection in spatial data sets are furnished here.

Algorithm - III: Correlative Logistic Correction Based Outlier Identification & Removal Algorithm (**CLC-OIR**)

Step - 1. Accept the Text set as T[] from the Spatial Dataset

Step - 2. For each T[i] from T[] set

a. Calculate the Threshold, TH[i] as Eq. 24

b. Calculate the correction factor, Theta[i] as Eq. 25

c. If T[i] > TH[i]

d. Then,

i. Mark the data item or Data Point as Outlier

ii. Replace the Outlier as T[i] = TH[i] ± Theta[i]

Step - 3. Report the final outlier removed dataset as T[]

They suggest either dimension error or the people come with a supply, although outliers can happen by chance in virtually and supply. At the prior event one wants to lose them use statistics which can be robust to outliers, any particular you must be careful in applying and then although at the latter instance that they signify which the supply comes with skewness. A reason for outliers can be that a mix of two distributions, which could signal identification versus dimension malfunction, or maybe just two, means of mix version model this really.

The results obtained from the proposed algorithms are highly satisfactory and are discussed in the next section of the work.

V. Results and Discussions

The proposed algorithm is evaluated on standard and benchmark dataset [14] and are highly satisfactory. In the section of the work, the obtained results are furnished and discussed.

Firstly, the image sets are introduced with some additional noises for better identification of the improvements over the traditional algorithms with the proposed algorithm. The initial noise induction results are elaborated here [Table I]. The noise types induced as Type-1 for Salt and Pepper Noise, Type-2 for Gaussian Noise, Type-3 for Sparkle Noise and finally Type-4 for passion Noise.

The noise levels are also visualized graphically here [Fig. 2].

Secondly, the results from the image noise reduction algorithm is furnished [Table II] and the step by step comparison is also done with two benchmark work by A. Erturk [11] and D. Cesmeci [12].

The results are also visualized graphically here [Fig. 3].

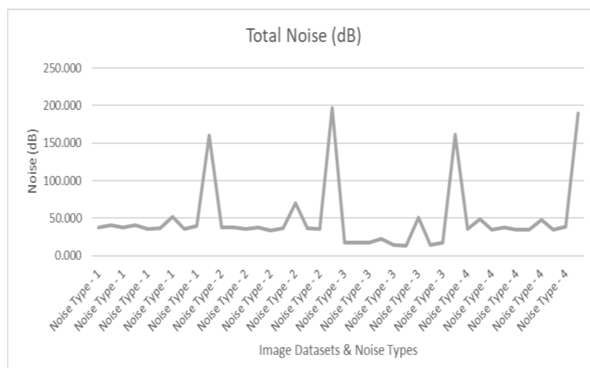


Fig. 2. Initial Noise Level Analysis.

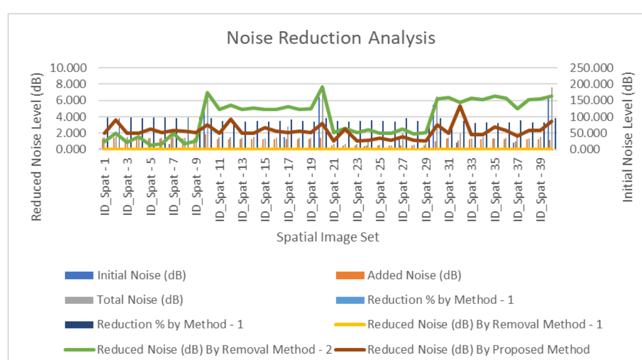


Fig. 3. Noise Level Reduction Analysis.

TABLE. I. NOISE INDUCTION IN SPATIAL IMAGES

	Data Item	Initial Noise (dB)	Added Noise (dB)	Total Noise (dB)
Noise Type - 1	ID_Spat - 1	4.250	33.860	38.110
Noise Type - 1	ID_Spat - 2	5.865	34.965	40.831
Noise Type - 1	ID_Spat - 3	4.038	33.689	37.726
Noise Type - 1	ID_Spat - 4	6.357	34.315	40.672
Noise Type - 1	ID_Spat - 5	1.775	33.648	35.422
Noise Type - 1	ID_Spat - 6	2.814	33.642	36.456
Noise Type - 1	ID_Spat - 7	16.436	35.586	52.021
Noise Type - 1	ID_Spat - 8	1.387	34.000	35.387
Noise Type - 1	ID_Spat - 9	5.829	33.903	39.732
Noise Type - 1	ID_Spat - 10	114.262	45.715	159.976
Noise Type - 2	ID_Spat - 11	5.950	31.544	37.494
Noise Type - 2	ID_Spat - 12	5.865	31.474	37.339
Noise Type - 2	ID_Spat - 13	4.038	31.568	35.605
Noise Type - 2	ID_Spat - 14	6.357	31.600	37.957
Noise Type - 2	ID_Spat - 15	1.420	31.699	33.119
Noise Type - 2	ID_Spat - 16	4.925	31.682	36.607
Noise Type - 2	ID_Spat - 17	38.350	32.114	70.464
Noise Type - 2	ID_Spat - 18	4.854	31.472	36.327
Noise Type - 2	ID_Spat - 19	3.886	31.556	35.441
Noise Type - 2	ID_Spat - 20	159.966	36.839	196.805
Noise Type - 3	ID_Spat - 21	5.100	12.183	17.283
Noise Type - 3	ID_Spat - 22	5.865	11.919	17.785
Noise Type - 3	ID_Spat - 23	4.845	12.080	16.925
Noise Type - 3	ID_Spat - 24	10.595	12.063	22.658
Noise Type - 3	ID_Spat - 25	2.485	12.253	14.737
Noise Type - 3	ID_Spat - 26	1.407	11.980	13.387
oise Type - 3	ID_Spat - 27	38.350	12.638	50.988
Noise Type - 3	ID_Spat - 28	2.774	11.853	14.627
Noise Type - 3	ID_Spat - 29	5.829	12.014	17.842
Noise Type - 3	ID_Spat - 30	137.114	24.586	161.700
Noise Type - 4	ID_Spat - 31	2.550	32.927	35.476
Noise Type - 4	ID_Spat - 32	20.529	27.979	48.507
Noise Type - 4	ID_Spat - 33	1.615	32.483	34.099
Noise Type - 4	ID_Spat - 34	6.357	31.017	37.374
Noise Type - 4	ID_Spat - 35	1.065	33.839	34.904
Noise Type - 4	ID_Spat - 36	2.814	32.195	35.009
Noise Type - 4	ID_Spat - 37	21.914	25.968	47.882
Noise Type - 4	ID_Spat - 38	3.467	31.406	34.874
Noise Type - 4	ID_Spat - 39	6.800	31.742	38.542
Noise Type - 4	ID_Spat - 40	159.966	29.772	189.738

TABLE. II. NOISE DETECTION AND REMOVAL FOR SPATIAL IMAGES

Data Item	Total Noise (dB)	Reduced Noise (dB) By Removal Method – 1 [11]	Reduced Noise (dB) By Removal Method – 2 [12]	Reduced Noise (dB) By Proposed Method
ID_Spat - 1	38.110	0.011	0.948	2.033
ID_Spat - 2	40.831	0.012	2.002	3.632
ID_Spat - 3	37.726	0.011	0.847	2.028
ID_Spat - 4	40.672	0.011	1.550	2.025
ID_Spat - 5	35.422	0.011	0.527	2.471
ID_Spat - 6	36.456	0.011	0.751	2.093
ID_Spat - 7	52.021	0.012	2.032	2.380
ID_Spat - 8	35.387	0.011	0.728	2.218
ID_Spat - 9	39.732	0.011	1.066	2.059
ID_Spat - 10	159.976	0.015	6.931	2.991
ID_Spat - 11	37.494	0.009	4.947	1.972
ID_Spat - 12	37.339	0.009	5.395	3.733
ID_Spat - 13	35.605	0.009	4.939	1.961
ID_Spat - 14	37.957	0.009	5.115	1.998
ID_Spat - 15	33.119	0.009	4.888	2.673
ID_Spat - 16	36.607	0.009	4.905	2.292
ID_Spat - 17	70.464	0.009	5.243	2.100
ID_Spat - 18	36.327	0.009	4.892	2.297
ID_Spat - 19	35.441	0.009	4.974	2.072
ID_Spat - 20	196.805	0.011	7.658	3.163
ID_Spat - 21	17.283	0.004	2.118	1.051
ID_Spat - 22	17.785	0.004	2.635	2.496
ID_Spat - 23	16.925	0.004	2.091	1.033
ID_Spat - 24	22.658	0.004	2.385	1.134
ID_Spat - 25	14.737	0.004	1.960	1.358
ID_Spat - 26	13.387	0.004	1.987	1.178
ID_Spat - 27	50.988	0.004	2.498	1.578
ID_Spat - 28	14.627	0.003	1.936	1.175
ID_Spat - 29	17.842	0.004	2.119	1.088
ID_Spat - 30	161.700	0.008	6.224	3.038
ID_Spat - 31	35.476	0.010	6.399	2.034
ID_Spat - 32	48.507	0.008	5.771	5.326
ID_Spat - 33	34.099	0.009	6.275	1.846
ID_Spat - 34	37.374	0.009	6.123	1.816
ID_Spat - 35	34.904	0.010	6.564	2.808
ID_Spat - 36	35.009	0.009	6.236	2.343
ID_Spat - 37	47.882	0.008	5.024	1.634
ID_Spat - 38	34.874	0.009	6.115	2.361
ID_Spat - 39	38.542	0.009	6.175	2.315
ID_Spat - 40	189.738	0.010	6.566	3.445

TABLE. III. NOISE REDUCTION PERCENTAGE ANALYSIS FOR SPATIAL IMAGES

Data Item	Reduction % by Method – 1	Reduction % by Method – 2	Reduction % by Proposed Method
ID_Spat - 1	88.819	86.362	83.515
ID_Spat - 2	85.606	80.731	76.740
ID_Spat - 3	89.267	87.052	83.921
ID_Spat - 4	84.342	80.559	79.392
ID_Spat - 5	94.958	93.503	88.013
ID_Spat - 6	92.250	90.220	86.538
ID_Spat - 7	68.384	64.501	63.831
ID_Spat – 8	96.049	94.023	89.814
ID_Spat – 9	85.302	82.648	80.148
ID_Spat – 10	28.567	24.243	26.706
ID_Spat – 11	84.107	70.938	78.871
ID_Spat – 12	84.267	69.843	74.294
ID_Spat – 13	88.634	74.787	83.151
ID_Spat – 14	83.228	69.777	77.987
ID_Spat – 15	95.685	80.954	87.642
ID_Spat – 16	86.521	73.148	80.286
ID_Spat – 17	45.562	38.136	42.595
ID_Spat – 18	86.612	73.171	80.314
ID_Spat – 19	89.010	75.001	83.191
ID_Spat – 20	18.713	14.827	17.111
ID_Spat – 21	70.472	58.236	64.415
ID_Spat – 22	66.999	52.206	52.989
ID_Spat – 23	71.350	59.019	65.269
ID_Spat – 24	53.224	42.712	48.235
ID_Spat – 25	83.116	69.840	73.926
ID_Spat – 26	89.463	74.644	80.693
ID_Spat – 27	24.779	19.887	21.692
ID_Spat – 28	81.013	67.798	73.000
ID_Spat – 29	67.313	55.456	61.235
ID_Spat – 30	15.200	11.356	13.326
ID_Spat – 31	92.786	74.776	87.078
ID_Spat – 32	57.662	45.782	46.699
ID_Spat – 33	95.236	76.860	89.848
ID_Spat – 34	82.967	66.608	78.131
ID_Spat – 35	96.921	78.143	88.905
ID_Spat – 36	91.935	74.150	85.269
ID_Spat – 37	54.217	43.740	50.821
ID_Spat – 38	90.032	72.524	83.287
ID_Spat – 39	82.333	66.336	76.349
ID_Spat – 40	15.686	12.231	13.875

As the reduced noise levels are concerns, much reductions can be observed by the first methods [XX]. However, the other image integrity-based factors are highly compromised. The details are furnished and discussed further in the sections of the work. Thirdly, the noise reduction percentage by these three methods are also analysed here [Table III].

The results are also visualized graphically here [Fig. 4].

Further, the image information loss analysis is carried out. One of the most prominent measure of the information loss from the images are size of the image apart from the pixel intensity and scale density. The size comparison after the noise reduction is furnished here [Table IV].

The results are also visualized graphically here [Fig. 5].

The results are also visualized graphically here [Fig. 6].

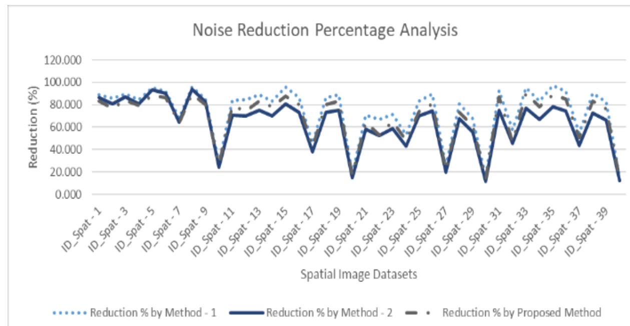


Fig. 4. Noise Level Reduction Percentage Analysis.

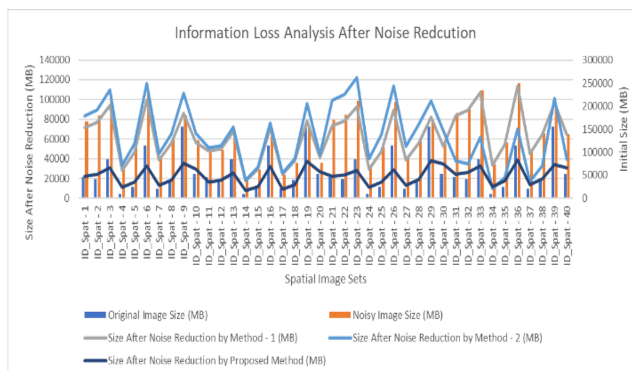


Fig. 5. Information Loss Percentage Analysis is Formulated [Table V].

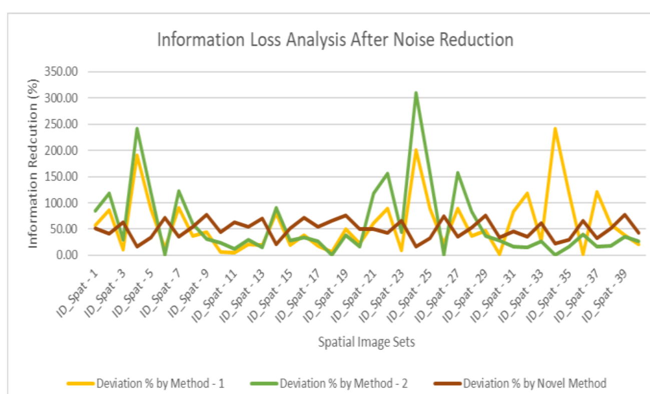


Fig. 6. Information Percentage Loss Analysis.

TABLE. IV. SIZE ANALYSIS AFTER THE NOISE REDUCTION FOR SPATIAL IMAGES

Data Item	Original Image Size (MB)	Noisy Image Size (MB)	Size After Noise Reduction by Method - 1 (MB)	Size After Noise Reduction by Method - 2 (MB)	Size After Noise Reduction by Proposed Method (MB)
ID_Spat - 1	45437	167331	71693	83712	22351
ID_Spat - 2	41130	178854	76705	89544	24447
ID_Spat - 3	84698	214678	94117	109470	30794
ID_Spat - 4	9755	66685	28429	33312	11351
ID_Spat - 5	24968	109963	46808	54912	16531
ID_Spat - 6	115202	228475	99408	116482	32575
ID_Spat - 7	20457	91513	38863	45497	13082
ID_Spat - 8	40564	128879	55745	64985	18351
ID_Spat - 9	155164	179606	85817	106752	35199
ID_Spat - 10	53071	125578	56409	66033	29640
ID_Spat - 11	45437	99249	47950	51110	16721
ID_Spat - 12	41130	104508	49877	53377	18594
ID_Spat - 13	84698	133089	67833	72264	25537
ID_Spat - 14	9755	38124	17532	18605	7775
ID_Spat - 15	24968	63250	29653	31839	12001
ID_Spat - 16	115202	144451	71066	76103	31859
ID_Spat - 17	20457	51769	24214	25793	9468
ID_Spat - 18	40564	76102	38241	40258	13885
ID_Spat - 19	155164	157122	78464	95699	37598
ID_Spat - 20	53071	77273	41105	44247	26842
ID_Spat - 21	45437	169599	73438	99399	22528
ID_Spat - 22	41130	180482	78018	105462	23567
ID_Spat - 23	84698	212141	92933	122483	28543
ID_Spat - 24	9755	67926	29310	39980	11391
ID_Spat - 25	24968	110082	47055	64150	16718
ID_Spat - 26	115202	208493	89502	113856	28939
ID_Spat - 27	20457	90795	38538	52554	13168
ID_Spat - 28	40564	128375	55702	74616	19211
ID_Spat - 29	155164	168397	82160	98588	38230
ID_Spat - 30	53071	117449	52395	68099	35066
ID_Spat - 31	45437	184563	83364	38048	24542
ID_Spat - 32	41130	197409	89594	35099	26516
ID_Spat - 33	84698	234286	106951	61869	32634
ID_Spat - 34	9755	73556	33317	9857	11937
ID_Spat - 35	24968	120899	54596	20914	17454
ID_Spat - 36	115202	249398	113641	70122	38501
ID_Spat - 37	20457	100838	45352	17157	13953
ID_Spat - 38	40564	141868	64307	33048	19995
ID_Spat - 39	155164	195594	96521	101160	33836
ID_Spat - 40	53071	138009	64324	39223	30793

TABLE. V. SIZE ANALYSIS AFTER THE NOISE REDUCTION FOR SPATIAL IMAGES

Data Item	Original Image Size (MB)	Noisy Image Size (MB)	Deviation % by Method – 1	Deviation % by Method – 2	Deviation % by Novel Method
ID_Spat - 1	45437	167331	57.79	84.24	50.81
ID_Spat - 2	41130	178854	86.49	117.71	40.56
ID_Spat - 3	84698	214678	11.12	29.25	63.64
ID_Spat - 4	9755	66685	191.43	241.49	16.36
ID_Spat - 5	24968	109963	87.47	119.93	33.79
ID_Spat - 6	115202	228475	13.71	1.11	71.72
ID_Spat - 7	20457	91513	89.97	122.40	36.05
ID_Spat - 8	40564	128879	37.42	60.20	54.76
ID_Spat - 9	155164	179606	44.69	31.20	77.31
ID_Spat - 10	53071	125578	6.29	24.42	44.15
ID_Spat - 11	45437	99249	5.53	12.49	63.20
ID_Spat - 12	41130	104508	21.27	29.78	54.79
ID_Spat - 13	84698	133089	19.91	14.68	69.85
ID_Spat - 14	9755	38124	79.72	90.72	20.30
ID_Spat - 15	24968	63250	18.76	27.52	51.93
ID_Spat - 16	115202	144451	38.31	33.94	72.35
ID_Spat - 17	20457	51769	18.37	26.08	53.72
ID_Spat - 18	40564	76102	5.73	0.75	65.77
ID_Spat - 19	155164	157122	49.43	38.32	75.77
ID_Spat - 20	53071	77273	22.55	16.63	49.42
ID_Spat - 21	45437	169599	61.63	118.76	50.42
ID_Spat - 22	41130	180482	89.69	156.41	42.70
ID_Spat - 23	84698	212141	9.72	44.61	66.30
ID_Spat - 24	9755	67926	200.46	309.84	16.77
ID_Spat - 25	24968	110082	88.46	156.93	33.04
ID_Spat - 26	115202	208493	22.31	1.17	74.88
ID_Spat - 27	20457	90795	88.39	156.90	35.63
ID_Spat - 28	40564	128375	37.32	83.95	52.64
ID_Spat - 29	155164	168397	47.05	36.46	75.36
ID_Spat - 30	53071	117449	1.27	28.32	33.93
ID_Spat - 31	45437	184563	83.47	16.26	45.99
ID_Spat - 32	41130	197409	117.83	14.66	35.53
ID_Spat - 33	84698	234286	26.27	26.95	61.47
ID_Spat - 34	9755	73556	241.54	1.05	22.37
ID_Spat - 35	24968	120899	118.66	16.24	30.09
ID_Spat - 36	115202	249398	1.36	39.13	66.58
ID_Spat - 37	20457	100838	121.69	16.13	31.79
ID_Spat - 38	40564	141868	58.53	18.53	50.71
ID_Spat - 39	155164	195594	37.79	34.80	78.19
ID_Spat - 40	53071	138009	21.20	26.09	41.98

Hence, it is natural to realize that, the information loss is the least by the proposed method. Further, as the spatial dataset does not only concern the image data, rather also the text data. Hence, this algorithm also analyses the missing value detection and outlier removal results.

The missing value analysis is carried out for the complete spatial dataset. However, only few furnished here [Table VI].

The results are also visualized graphically here [Fig. 7].

The outlier analysis is carried out for the complete spatial dataset. However, only few are furnished here [Table VII].

TABLE. VI. MISSING VALUE DETECTION ANALYSIS

Data Item	Number of Data Items	Number of Missing Values	Detected Missing Values	Accuracy (%)
ID_Spat - 1	400	8	7	87.50
ID_Spat - 2	393	9	9	100.00
ID_Spat - 3	387	7	7	100.00
ID_Spat - 4	310	5	5	100.00
ID_Spat - 5	320	5	5	100.00
ID_Spat - 6	337	5	4	80.00
ID_Spat - 7	301	10	6	60.00
ID_Spat - 8	369	7	6	85.71
ID_Spat - 9	322	9	7	77.78
ID_Spat - 10	329	7	7	100.00

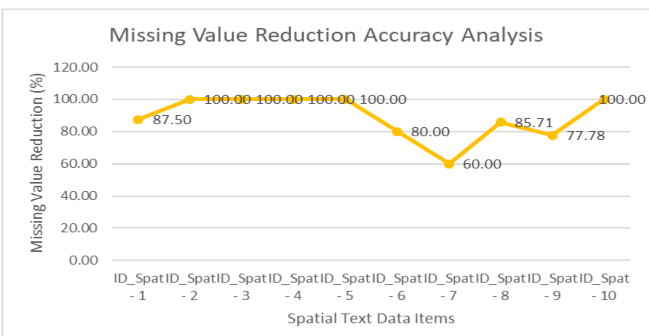


Fig. 7. Missing Value Replacement Accuracy Analysis.

TABLE. VII. OUTLIER VALUE DETECTION ANALYSIS

Data Item	Number of Data Items	Number of Missing Values	Detected Missing Values	Accuracy (%)
ID_Spat - 1	400	4	4	100.00
ID_Spat - 2	393	8	6	75.00
ID_Spat - 3	387	7	5	71.43
ID_Spat - 4	310	4	4	100.00
ID_Spat - 5	320	7	7	100.00
ID_Spat - 6	337	6	5	83.33
ID_Spat - 7	301	5	5	100.00
ID_Spat - 8	369	5	4	80.00
ID_Spat - 9	322	4	4	100.00
ID_Spat - 10	329	7	6	85.71

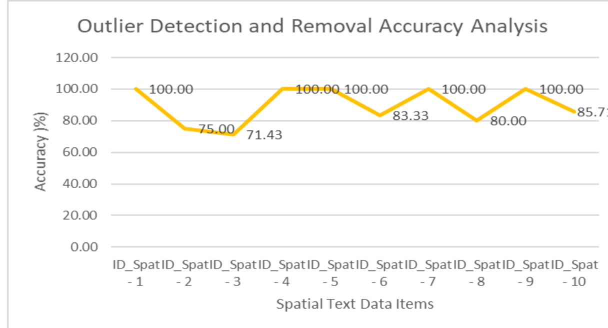


Fig. 8. Outlier Detection and Replacement Accuracy Analysis.

The results are also visualized graphically here [Fig. 8].

Henceforth, with the detailed analysis of the results obtained from the proposed algorithms, in the next section of this work, the comparative analysis is furnished.

VI. COMPARATIVE ANALYSIS

Although, the step by step comparisons are carried out with highly benchmarked parallel research outcomes, in the previous section of this work, the summarized comparative analysis is carried out here [Table VIII].

The improvements over the existing algorithms are notable and the reasons are elaborated here.

Firstly, instead quite a few complexes mean of plasma normalization are constructing on arrangement preserving transformations homeomorphisms and also diffeomorphisms given that they take sleek sub manifolds effortlessly throughout conversion. Diffeomorphisms are created inside today’s area of computational anatomy predicted on diffeomorphic leaks, additionally referred to as diffeomorphic mapping. But such transformations by way of diffeomorphic aren’t additive, even though they produce a set with work article and behaving non-linearly to the graphics by the way of team actions.

Secondly, any process for spectral imaging, allowing qualitative and qualitative characterization of their air and also of this outside. These dimensions may be properly utilized for atmospheric gases, the dimension of these concentrations and identifications of outside substances and unambiguous direct, then the mission of their participation of signs that were blended.

TABLE. VIII. SUMMARIZED COMPARATIVE ANALYSIS

Method Name	Noise Reduction Percentage Mean (%)	Image Information Loss Percentage Mean (%)	Missing Value Detection and Reduction Accuracy Mean (%)	Outlier Detection and Reduction Accuracy Mean (%)	Model Complexity
A. Ertürk et al [11]	73.96	59.51	59.21	57.55	High
D. Çeşmeci et al. [12]	63.67	60.65	61.75	62.62	Moderate
Proposed Method	68.01	50.31	89.10	89.55	Low

Third, in machine-learning and predictive optimization, loss works for classifications are all computationally viable loss purposes representing the cost covered inaccuracy of predictions from classification issues as issues of identifying that category a specific monitoring belongs to.

Finally, there will be to unmixing a method always to undo the procedure that is blending. Ordinarily, of blending 2 types are all supposed: nonlinear and linear. Vertical blending like being level models that the earth and episode sun onto the earth results in the substance to digitize a certain sum of their episode energy right back into the detector. Just about every pixel is described as a sum of the vitality slabs of substances. Just about every material contributes to the monitoring of this sensor within a mode that is favourable. A conservation of electricity restriction is detected inducing this mixture's weights to amount to a in addition to having the favourable.

Henceforth, after the detailed discussion on the comparative analysis, this work presence the final research conclusion in the next section of the work.

VII. Conclusion

Automation in the road accident prediction is highly dependent on the road conditions. Thus, correct detections of the road conditions are one of the most important aspects. The road data or the spatial road information is highly prune to the noise. Failing to justify the de-noising process of the spatial data with respect to the image intensity or the information loss from the parallel research outcomes, this work proposes a novel adaptive moments-based image de-noising methods with the use of adaptive intensity calculations for image segments. Further in order to also de-noise the endmember data items, extracted from the spatial data, this work introduces two other methods as adaptive logistic estimation and corrective logistic estimation for the de-noising purpose. The work results in nearly 90% accuracy of de-nosing process. This work can be considered as a newer benchmark for de-noising the spatial road data for making the further processing highly accurate.

References

[1] D. A. Landgrebe. "Signal Theory Method in Multispectral Remote Sensing",Hoboken, NJ, USA:Wiley.2003.

[2] J. M. Bioucas-Dias. Hyperspectral unmixing overview: Geometrical statistical and sparse regression-based approaches", IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 2012 vol.,no.2,pp.354-379.

[3] N. Keshava J. F. Mustard. "Spectral unmixing", IEEE Signal Process. Mag. 2002 vol. 19, no. 1, pp. 44-57, Jan.

[4] M. E. Winter. "N-FINDR: An algorithm for fast autonomous spectral endmember determination in hyperspectral data", P roc. SPIE,1999. vol. 3753, pp. 266-277.

[5] J.M.P.Nascimento,,J.M.Bioucas-Dias.-"Vertex-component, Analysis: A fast algorithm tounmi- hyperspectral data ", IEEE Trans.Geosci.Remote Sens. 2005 vol. 43, no. 4, pp. 898-910, Apr.

[6] J. Li, J. Bioucas-Dias. Minimum volume simplex Analysis: A fast algorithm to unmix hyperspectral Data ,Proc. IEEE Int. Geosci. Remote Symp.2008 (IGARSS),vol.3,pp.250-253.

[7] Alp Ertürk. Enhanced Unmixing-Based Hyperspectral Image Denoising Using Spatial Preprocessing, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Volume: 8, Issue: 6. 2015.

[8] M. Zortea, A. Plaza. Spatial preprocessing for endmember extraction, IEEE Trans. Geosci.RemoteSens.,vol.47,no.8,pp.2679-2693.2009.

[9] G. Martin, A. Plaza. Region-based Spatial preprocessing for endmember extraction and spectral unmixing. IEEE Trans. Geosci. Remote Sens., vol.8, no. 4,pp.745-749. 2011.

[10] G. Martin, A. Plaza. Spatial-spectral preprocessing prior to endmember identification and unmixing of remotely sensed hyperspectral data, 2012. IEEE-J.Sel. opics Appl. Earthbserv.RemoteSens.,vol.5,no.2,pp.380395.

[11] Ertürk, D. Çeşmeci, M. K. Güllü, D.Gerçek, and S. Ertürk. "Integrating Anomaly detection to spatial preprocessing for endmember extraction of Hyperspectralimages".2013,Proc.IEEE,Geosci.RemoteSens .Symp.(IGARSS),pp.1087-1090.

[12] A.Ertürk, D. Çeşmeci. M. K. Güllü D. Gerçek, and S. Ertürk. "Endmember extraction guided by anomalies and homogeneous regions forhyperspectralimages",2014,IEEE.Sel.opicsAppl.EarthObserv.Remote Sens.,vol.7,no.8,pp.30-3639,Aug.

[13] C. Li, L. Wang. Incorporating spatial information in spectral unmixing: A review .RemoteSens.Environ.vol.149,pp.70-87,Jun.2015.

[14] Dataset: Indian Surface Maintenance Authority Image Samples. D. A. Landgrebe. "Signal Theory Method in Multispectral Remote Sensing",Hoboken, NJ, USA:Wiley.2003.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

Blockchain-based Electronic Voting System with Special Ballot and Block Structures that Complies with Indonesian Principle of Voting

Gottfried Christophorus Prasetyadi[1]
Graduate Program on Information System
Gunadarma University, Depok 16424, Indonesia

Achmad Benny Mutiara[2], Rina Refianti[3]
Faculty of Computer Science and Information Technology
Gunadarma University, Depok 16424, Indonesia

***Abstract*—Blockchain technology could be implemented not only in digital currency, but also in other fields. One such implementation is in democratic life, namely voting. This research focuses on designing a blockchain-based electronic voting system for medium to large-scale usage that complies with law, specifically voting principles in Indonesia. In this research, we proposed the following: a ballot design as block transaction employing UUID version 4, a modified block structure using SHA3-256 hash algorithm, and a voting protocol. The minimum length of a ballot is 43 bytes (excluding ECDSA signature) if one character is used as candidate's identifier and timestamp is stored as integer. We built a simulation program using Python-based Django web framework to cast 10,000 votes and mine them into blocks. Tampered transactions in each block could be detected and restored by synchronizing data with another node. We also evaluated the proposed system. By using this system, voters can exercise voting principles in Indonesia: direct, public, free, confidential, honest, and fair.**

Keywords—Blockchain; voting; design; simulation; Python

I. Introduction

Information and communication technology is advancing rapidly. The performance and efficiency of Central Processing Unit (CPU) as the heart of a computer have continued to improve in the last few decades. Moore's Law, based on Gordon Moore's observation in 1965 and later adjustment in 1975, stated that the size of transistors were shrinking so fast that every two years, twice as many could fit onto a single computer chip [1]. This advancement has revolutionized many aspects in our social life and government. One such case that is going to be discussed in this study is voting.

Democratic countries, such as the Republic of Indonesia, guarantee the rights of their citizens to participate in decision-making, for example, to choose leaders by the mean of voting. By definition, voting (to vote) is "a formal indication of a choice between two or more candidates or courses of action, expressed typically through a ballot or a show of hands or by voice". In Indonesia, this right is listed on the state's constitution, namely Undang-undang Dasar Negara Republik Indonesia 1945 (UUD NRI 1945) article 28J paragraph 3: "everyone has the right to freedom of association, assembly, and issuing opinions".

Nowadays, voting process may be done electronically. Several electronic voting systems had been developed such as VOTAN (Votes Analyzer) for conducting electronic elections through the Internet securely. It is ideal for small communities such as organizations, universities and chambers [2]. It uses a centralized database, just like many other similar systems.

Centralized systems have common weaknesses. The data are stored centrally, so they have central point of failure, which can be exploited by computer crackers. Those systems are usually handled by single organization, so the data can be manipulated secretly by those who have administrative access to the database [3]. The recent development of blockchain technology can solve this problem.

The first work on cryptographically secured chain of blocks was published in 1991 in order to implement a system where documents' timestamps could not be modified [4]. In 2008, Satoshi Nakamoto, whose identity is still unknown, wrote about a "purely peer-to-peer version of electronic cash" known as Bitcoin [5]. Since then, blockchain made its public debut. Over time, people started to realize that blockchain could be used beyond cryptocurrency and they started to explore how blockchain could enhance many existing systems, including in voting process.

This study focuses on the design of several important components of the blockchain-based electronic voting system, and discusses the implementation of the proposed system for secure electronic voting to guarantee the rights of people, especially Indonesian citizens. The proposed system must follow the rules and principles recognized by the state.

This study is limited by the following. First, voters have to able to identify themselves using pseudonym. Second, the proposed system is intended for medium to large-scale usage, not small-scale (which often does not require costly effort). Third, node registration and public key storage are not discussed. Fourth, the simulation and testing are done on the local machine. Fifth and last, this study does not cover the solution for disabled people to access the system.

This paper is organized as follows. Section II contains comprehensive theoretical bases and proposed methods. Section III contains testing results and discussion. Section IV contains concluding remarks and possibilities of further improvements.

II. Related Works

Author in [3] proposed a blockchain-based system to record the results of elections using predetermined turn (flowchart shown in Fig. 1), instead of proof of work. Each node represented one voting place (called TPS/Tempat Pemungutan Suara) that produced one block containing one transaction comprising the sum of votes for each candidate. The researcher managed to generate 500,000 valid nodes in the simulation. However, it was not possible to identify and verify individual ballot, which this study tries to provide. Also, from their method, the predetermined turn system allowed only selected node (government-owned) to form the network.

Author in [7] proposed a protocol that employed blind digital signature and formulation of valid vote message, shown in Fig. 2. Their analysis did not involve simulation to prove their method, although they claim that the protocol can be implemented easily into Bitcoin network.

Author in [8] proposed a decentralized e-voting system that used seven roles to guarantee public and transparent voting process while ensuring voter's anonymity. Those roles are voters, registration server, authentication server, voting website, recording center, distributed data servers, and smart contract. Their proposed design involved Paillier homomorphic encryption, which we do not use as there is no arithmetic operation on cipher data in our solution.

Fig. 1. Flowchart from [6].

$$\overbrace{\underbrace{10}_{2}\,\underbrace{0000\cdots0000}_{y}\;\underbrace{0110\cdots1110}_{z\text{ bits random string}}}^{n}$$

Fig. 2. Ballot Structure from [7] with Two First Characters used to Identify each Candidate.

III. The Material and Method

A. Electronic Voting and Election Law

Electronic voting refers to voting process that utilizes electronic devices and other modern technologies to cast and count the votes. Electronic voting can be held via internet, which the voters submit their votes to the voting organizer, from any location [9]. Organizers must employ any means necessary to ensure authentication and authorization for every cast ballot.

Specifically in Indonesia, Law (*Undang-undang*) number 7 year 2017 states in Article 2 that general election must comply with the following principles:

1) Direct: Each voter must cast his/her vote directly and not represented by other person or party.

2) Public: Every eligible member of society may participate in the voting, to cast his/her vote.

3) Free: A voter chooses candidate by his own will, not under threat or forced.

4) Confidential: Only the respective voter knows a voter's choice.

5) Honest: Every election and voting must comply with the regulation to guarantee the right of the voters, and that each vote cast has the same value.

6) Fair: All voters have equal right to vote, without any special privilege or discrimination.

Those principles formally applies to national election (such as electing president or regional representatives), although there is no reason not to use it as basis for any other type of voting in a democratic country such as Republic of Indonesia.

B. Blockchain

Blockchain is a shared ledger of transactions. The transactions are ordered and grouped into blocks. Currently, the real-world model is based on private databases that each organization maintains, whereas the distributed ledger can serve as a single source of truth for all member organizations that are using the blockchain. Blockchain is also a data structure, a linked list that uses hash pointers instead of normal pointers. Hash pointers are used to point to the previous block [10].

Bitcoin cryptocurrency with chain of blocks as its basis was proposed by [5]. Blockchain employs consensus algorithm to achieve decentralization of control. Consensus provides a way for all peers to agree and accept a single version of truth on the blockchain network. Bitcoin itself uses proof-of-work consensus to prove that enough computational resources have been spent before proposing a truth to be accepted by peers, therefore solving the double spending problem and Byzantine General's problem.

C. Secure Hash Algorithm 3 (SHA-3)

SHA-3 is a latest member of secure hash algorithm standards. A cryptographic hash function is a one-way function that uses mathematical algorithm to map data of any size (message) to a fixed size bit string (hash). SHA-3 is meant to be an alternative to SHA-2, after successful attacks were proven on MD5 and SHA-1. SHA-3 uses Keccak algorithm. It is based on un-keyed permutations as opposed to other usual hash functions' constructions that used keyed permutations. A new approach called sponge and squeeze construction is used in Keccak, which is a random permutation model. The draft of SHA-3 (FIPS 202) was approved on 2015 by US National Institute of Standards and Technology [11]. SHA-3 is considered safe against quantum attack [12]. The performance is in par with SHA-2 [13].

The variant used in this study is SHA-3 with 256-bit of output (SHA3-256).

D. Universally Unique Identifier (UUID) Version 4

A UUID is 128 bits value that is used to identify a piece of data or information in computer systems. Every UUID is unique. The uniqueness of each value is guaranteed when it is generated using standard methods, and it does not depend on the parties that generate it. The protocol to generate UUID is specified in RFC 4122 [14].

UUID version 4 (UUID4) is generated randomly, not time-based or name-based like previous versions of UUID. Its probability of collision is so small that it can be safely ignored.

It leaves 122 of its 128 bits available for random data. The probability to find a duplicate within 103 trillion UUID4s is one in a billion.

E. Elliptic Curve Digital Signature Algorithm (ECDSA)

Digital signature is mathematical scheme for authenticating digital data and documents. If a signed data has valid digital signature, then the recipient could safely believe that it was created by a known sender (authenticity), which the sender cannot deny (non-repudiation), and that the data is intact and not altered (integrity). Digital signature employs asymmetric cryptography, which means two distinct keys are needed.

The Elliptic Curve Digital Signature Algorithm (ECDSA) is the elliptic curve analogue of the Digital Signature Algorithm. The ECDSA is included in several standards, such as IEEE 1363-2000, ISO/IEC 15946-2, and FIPS PUB 186-4 (NIST). It is included in the cipher suites of the Transport Layer Security (TLS) protocol (RFC 4492) [15]. The elliptic curve is simply the set of points described by the equation (1) called Weierstrass normal form [16].

$$y^2 = x^3 + ax + b, \tag{1}$$

where $4a^3 + 27b^2 \neq 0$ to exclude singular curves.

ECDSA offers smaller key size than that of RSA-based ones for the same security level, allowing faster verification. Table I shows the comparison of RSA and ECC key sizes, while Table II shows the performance differences, which we measured by generating 1,000 signatures per algorithm to sign and verify 128 bytes message. The variant used in this study is ECDSA 256.

F. Methods

The right to vote is guaranteed by law in Indonesia. Voting process must follow voting principles: direct, public, free, confidential, honest, and fair. This study is aimed to provide a way for a voter to know whether his/her vote is recorded as-is, not just the summary of counts like in [6]. The proposed design will not implement Paillier cryptosystem, unlike [8]. The system will also provide a way to examine the counting process. Only valid voters may cast a vote.

The size of a transaction must be kept minimum, and the structure of the ballot must be simple and easy to understand. A block should contain as many transactions as possible, just like [5].

Blockchain is a distributed ledger. To ensure the accuracy and integrity of every record in a block, it must be sealed (mined) and chained with previous block. The sealing hash output is used as proof of work. Mining process requires effort, so the proposed system is intended for medium to large-scale usage.

In pre-voting phase, each potential voter generates UUID4 as pseudonym, a pair of public and private keys, and prepares legal documents. He/she then proceeds to validate his/her identity to the organizer, submit the public key, and keep his/her pseudonym and private key secret. It is up to the voting organizer to determine the best way to accomplish this. Fig. 3 shows the diagram of this phase.

TABLE. I. COMPARISON OF RSA AND ECC KEY SIZE [17]

RSA (bits)	ECC (bits)	Security Level (bits)
1024	160	80
2048	224	112
3072	256	128
7680	384	192
15,360	521	256

TABLE. II. PERFORMANCE DIFFERENCE BETWEEN ECDSA AND RSA-PSS

Component	ECDSA 256	RSA-PSS 3072
Avg. generation time (s)	0.085	1.161
Avg. signing time (s)	0.084	0.036
Avg. verification time (s)	0.169	0.002
Total time needed (s)	339.559	1,324.626

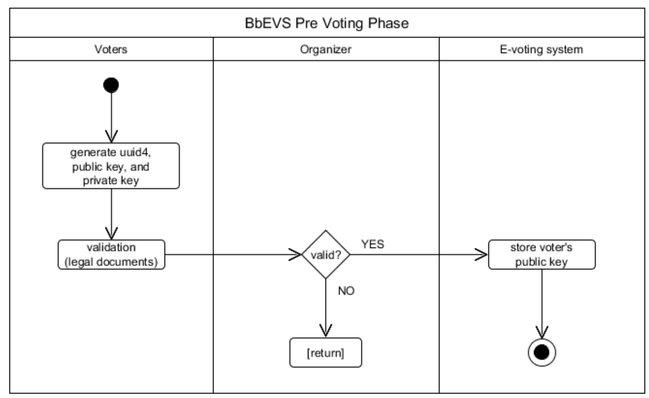


Fig. 3. Diagram of pre-voting phase

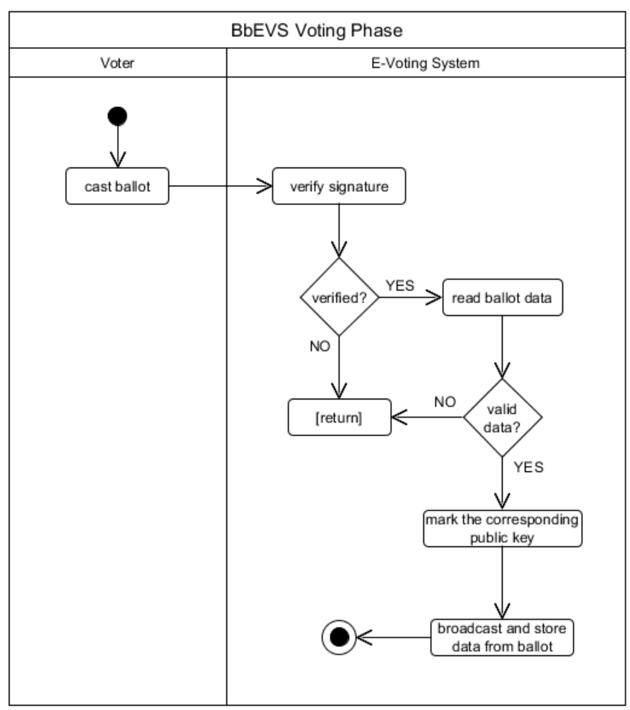


Fig. 4. Diagram of vote casting phase

Fig. 4 shows the diagram of vote casting phase. In this phase, each voter casts his/her own ballot after signing it. Once accepted by server, the ballot will be verified for authenticity and integrity before being relayed to all nodes. Data from a verified ballot are considered valid if the pseudonym is unique, candidate identifier is valid, and (optionally) the timestamp is considered reasonable.

Now we discuss the recording and counting phase. Ideally, a block is created when certain numbers of transactions (ballots) have been relayed to all nodes, and then that block is broadcasted. Finally, the voting result can be counted. The counted votes $V_{counted}$ should be less than or equal to the total votes cast, shown in (2).

$$V_{\text{counted}} = V_{\text{total}} - V_{\text{unmarked}} \tag{2}$$

The vote is 'unmarked' if the corresponding public key is never used for verification, or the data in the ballot are invalid.

In the proposed system, the ballot has the following structure:

$b_{\text{v\_id}} + b_{\text{c\_id}} + \text{t}$,

where b_{v_id} is the UUID as voter's ID (32 bytes), b_{c_id} is the candidate ID (length may vary), and t as timestamp (can be either integer or float value). Timestamp value may be either the ballot creation time or the time the ballot is received by electronic voting system. Thus, the minimum length of a ballot is 43 bytes. One or more fields may be added or modified depending on voting requirements. The following is an example of valid ballot:

ae19033a1d9a4f6cbaed53c6d2de1f730011540300734.584385.

As comparison, the size of a Bitcoin transaction is approximately 267 bytes. The structure of the block used in this study does not contain block version and difficulty target.

To study and analyze how the proposed system works, we created a simulation program. The software is a web-based application with the following details:

- Programming language: Python version 3.6.0,
- Framework: Python-based Django framework version 2.1.2,
- Database engine: SQLite 3,
- GUI: web-based (browser) (debug message and other details are printed on the system's shell),
- Server: Django integrated development server.

In this simulation, transactions are broadcasted to two nodes. One of the nodes acts as an always-honest node so all blocks and transactions can be compared later. The number of transactions, transactions per block, and puzzle difficulty can be adjusted to compensate the performance of the computer that runs the simulation.

The simulation comprises two sections: "block" and "chain" (Fig. 5).

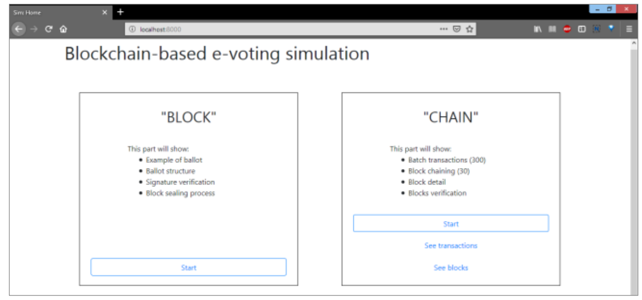


Fig. 5. Front Page of the Simulation Program.

1) "Block": This section shows the example of ballot and demonstrates signature verification and mining processes.

2) "Chain": This section demonstrates the batch generation of transactions, sealing (mining) process, detail of each block, verification, and data synchronization process.

The Python code to generate valid block hash is shown below:

```
# Try to seal the block and generate valid hash
nonce = 0
timestamp = datetime.datetime.now().timestamp()
while True:
    block_hash = SHA3_256.new(prev_hash +
merkle_h + nonce + timestamp).hexdigest()
    # Check whether this hash satisfies puzzle
requirement
    if block_hash[:pcount] == puzzle:
        break
    nonce += 1
```

The `prev_hash`, `merkle_h`, `nonce`, and `timestamp` are all encoded in binary so they are concatenated using '+' operator. `pcount` and `puzzle` are defined in separate configuration file.

The Python code to generate a single transaction with valid, random values is shown below:

```
# generate random, valid values
v_id = str(uuid4())
v_cand = _get_vote()
v_timestamp = _get_timestamp()
# directly fill the values and the block id for
simulation purpose
new_vote = Vote(id=v_id, vote=v_cand,
timestamp=v_timestamp)
new_backup_vote = VoteBackup(id=v_id,
vote=v_cand, timestamp=v_timestamp)
# "Broadcast" to two nodes
new_vote.save()
new_backup_vote.save()
print("#{} new vote: {}".format(i, new_vote)) #
debug
```

A database in each node is used to record and synchronize all transactions (ballots) and blocks. Table III and Table IV show the model structure of ballot and block, respectively. Fig. 6 shows their relationship.

TABLE. III. STRUCTURE OF VOTE (BALLOT) MODEL

Field	Type	Description
id	char	UUID as primary key
vote	int	Candidate id
timestamp	float	
block_id	int	Foreign key

TABLE. IV. STRUCTURE OF BLOCK MODEL

Field	Type	Description
id	int	Auto-increment, used for simulation
prev_h	char	Previous block hash
merkle_h	char	Merkle root hash
h	char	Block hash
nonce	int	An arbitrary number
timestamp	float	Block creation timestamp

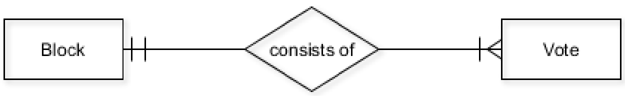


Fig. 6. Entity Relationship Diagram.

Some tests must be run to ensure the proposed system meets the following requirements. First, each user can examine their cast ballots after voting is over. Second, users can examine the detail of each block (hashes, nonce, number of transactions it contains, total number of blocks, etc.). Third, in case a node gets corrupted, it must be able to sync with majority of nodes aka "the agreed truth". A reasonably great number of dummy, valid ballots must be generated to run this test, i.e., 10,000. In our study, they are generated programmatically. The built-in user interface, i.e., web UI, is used to confirm the result.

The simulation was run and benchmarked on the computer with the following specifications:

- CPU: 4 Cores, up to 3.6 GHz,
- RAM: 8 (2x4) GB, 1,600 MHz,
- Hard drive: 466 GB capacity, Read 74.45 MB/s, Write 64.18 MB/s,
- Operating system: Windows 8.1 Pro, 64-bit.

IV. RESULTS AND DISCUSSION

A. Results

In the "Block" section of the simulation program, we cast a signed ballot using the web user interface (Fig. 7) as a single transaction and then sealed it into a block. After several trials, the block was mined successfully after a valid hash had been generated. In our test, the nonce was 71,252 and the puzzle difficulty required hash with four leading zeros. It took approximately 7.166 seconds (Fig. 8).

In the "Chain" section, we generated 10,000 votes (Fig. 9) in approximately 2,585 seconds (43 minutes 5 seconds), and broadcasted them to transaction pool. Each block comprised 20 votes as transactions. Thus, the maximum size of each block was 1,000 bytes. This size was decided to make sure that each block was small enough so all transactions in the block could be verified quickly. Finally, 500 blocks were created successfully; each one correctly contained 20 transactions. In this round, candidate #2 won by 3,416 votes (Fig. 10).

We tampered some records (transactions) on the database using a database management tool. All blocks and the transactions were then checked for integrity, and the system successfully detected blocks on the main node with tampered data, shown in Fig. 11. The troubled node had to synchronize its data with the majority of nodes on the network.

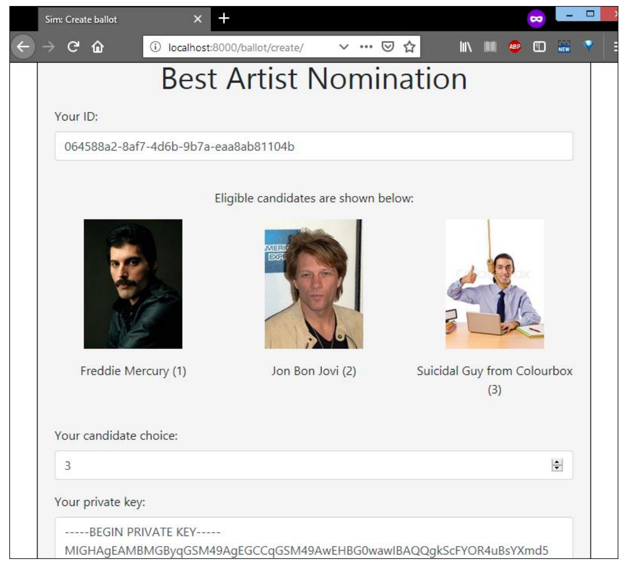


Fig. 7. A Ballot Example.

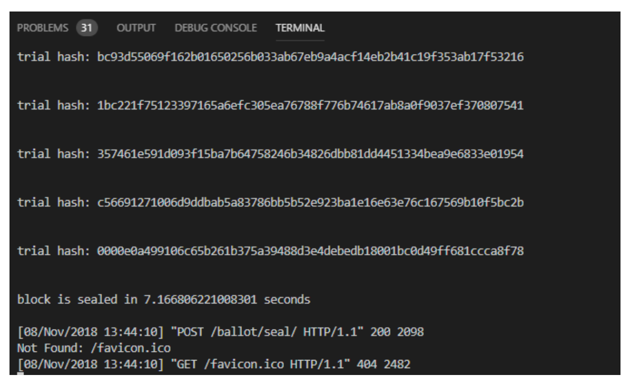


Fig. 8. Terminal Output Showing Mining Process.

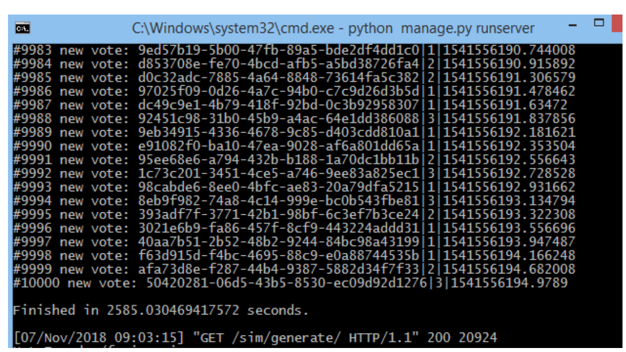


Fig. 9. Successfully Generated 10,000 Votes.

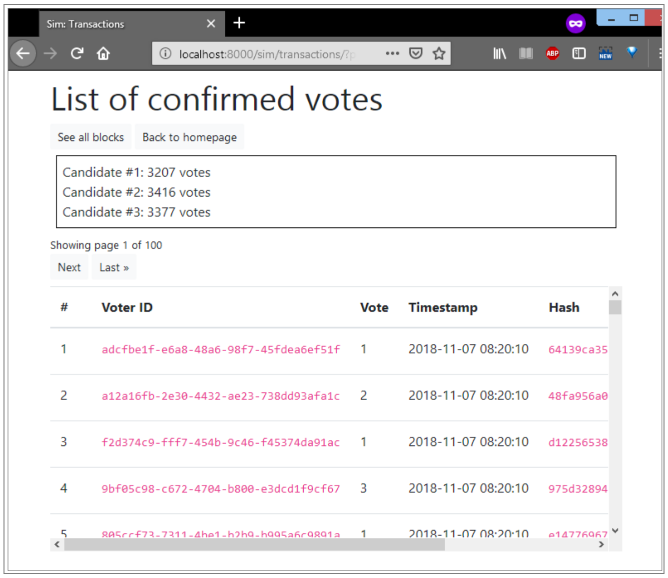


Fig. 10. Voting Result in the Simulation.

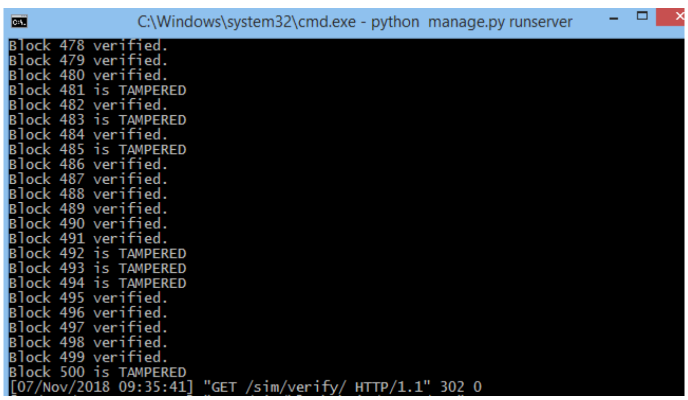


Fig. 11. The System Detected Blocks with Tampered Transactions.

We then synchronized the blocks in the main node by comparing them with the second (always-honest) node in the simulation. All the data were successfully restored, shown in Fig. 12 (for block #1). Fig. 13 shows block #1, including its header, status, and transactions (votes).

In our simulation, the timestamps were stored as float and each candidate was identified by only one character. The average ballot size was 47.42 bytes. Fig. 14 shows the approximated size of database for up to 9,999 transactions. That many transactions should require 474.152 KB of database.

B. Discussion

Each voter has to submit a public key after verification involving legal documents, so this system does not need fingerprint schema for digital signature. In the proposed system, after a signed ballot has been proven valid, the corresponding public key is then marked or deleted. This way, a voter could only vote once.

For the pseudonym, hash digest may also be used to replace UUID. The public key portion of ECDSA could possibly be used as the hash input. The reason of using UUID instead of hash digest in this study is that some hash algorithms that produce reasonably small digest output are prone to collision (such as MD5 and SHA1), and trimming long output could mean wasting computational power. Each UUID is unique and exposed only when a vote has been cast and verified. If somehow there are UUIDs with the same value in the pool, the first to be proven valid is the counted one (though this will never happen, if the protocol has been implemented correctly).

To ensure that a voter's location cannot be tracked, he/she can send the ballot via proxy, such as Tor. Inputting private key into web form to sign the vote is not recommended method; the voter should generate the valid ballot by his/her own and then send it via secure API. Due to the design of the proposed ballot, Paillier homomorphic encryption is not needed, contrary to the research by [8].

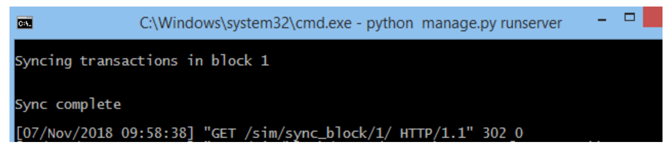


Fig. 12. All Transactions in a Block had been Synchronized.

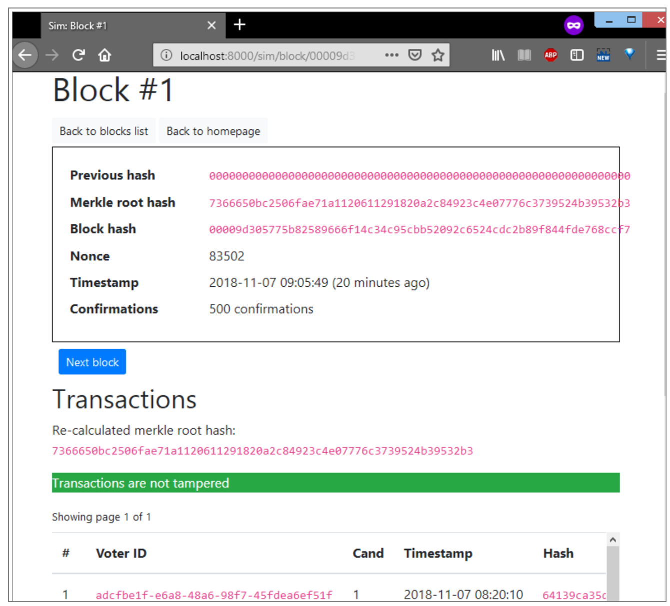


Fig. 13. Details of a Block.

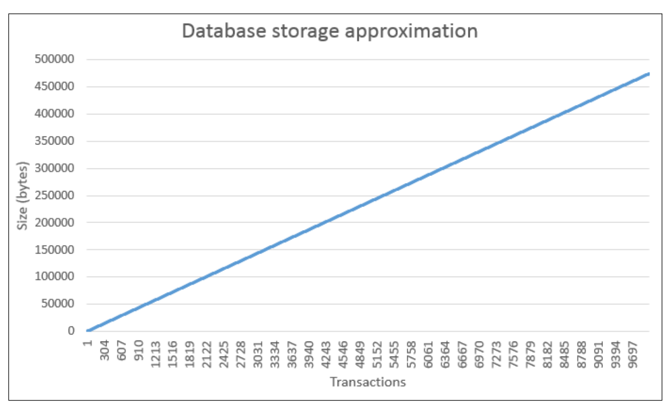


Fig. 14. Line Chart of Database Capacity Measurement.

V. Conclusions

Based on the simulation and testing results, the proposed system worked well and complied with Indonesian voting principles. Received ballots could be checked for authenticity and integrity using ECDSA, and then all the data in each ballot were checked for validity. Only valid ballots could be recorded as transactions, which were then mined into blocks. If a node had one or more tampered transactions, they could be detected and restored by comparing the data with other node. The usage of SHA-3 and ECDSA (instead of RSA signature) were meant to speed up the process of generating key pair and keep overall data size to minimum. In the future, the system could be tested inter-device on the local network or internet using secure API.

Acknowledgment

We would like to thank the Gunadarma Foundation for financial support of this research.

References

[1] J. L. Hennessy and D. A. Patterson, Computer Architecture: A Quantitative Approach, Burlington: Morgan Kaufmann, 2017.

[2] S. Valsamidis, S. Kontogiannis, T. Theodosiou and I. Petasakis, "A Web e-voting system with a data analysis component," Journal of Systems and Information Technology, vol. 20, no. 1, pp. 33-53, 2018.

[3] The Economist, "The great chain of being sure about things," 31 October 2015. [Online]. Available: https://www.economist.com/briefing/2015 /10/31/the-great-chain-of-being-sure-about-things. [Accessed 8 October 2018].

[4] A. Narayanan, J. Bonneau, E. Felten, A. Miller and S. Goldfeder, Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction, Princeton: Princeton University Press, 2016.

[5] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[6] R. Hanifatunnisa, "Design and Implementation of Blockchain Based E-Voting Recording System," Master's Program Thesis, Institut Teknologi Bandung, Bandung, 2017.

[7] Y. Liu and Q. Wang, "An e-voting protocol based on blockchain," IACR Cryptol. ePrint Arch., vol. 1043, p. 2017, 2017.

[8] J. Hsiao, R. Tso, C. Chen and M. Wu, "Decentralized E-Voting Systems Based on the Blockchain Technology," in Advances in Computer Science and Ubiquitous Computing, Singapore, Springer, 2017, pp. 305-309.

[9] D. Zissis and D. Lekkas, "Securing e-Government and e-Voting with an open cloud computing architecture," Government Information Quarterly, vol. 28, no. 2, pp. 239-251, 2011.

[10] I. Bashir, Mastering Blockchain, Birmingham: Packt Publishing Ltd., 2017.

[11] NIST, "Announcing Approval of Federal Information Processing Standard (FIPS) 202, SHA-3 Standard: Permutation-Based Hash and Extendable-Output Functions, and Revision of the Applicability Clause of FIPS 180-4, Secure Hash Standard," 5 August 2015. [Online]. Available: https://www.gpo.gov/fdsys/pkg/FR-2015-08-05/pdf/2015-19181.pdf. [Accessed 26 October 2018].

[12] M. Amy, O. Di Matteo, V. Gheorghiu, M. Mosca, A. Parent and J. Schanck, "Estimating the cost of generic quantum pre-image attacks on SHA-2 and SHA-3," in International Conference on Selected Areas in Cryptography, Cham, 2016.

[13] G. Bertoni, J. Daemen, M. Peeters, G. Assche and R. Keer, "Is SHA-3 slow?," 12 June 2017. [Online]. Available: https://keccak.team/2017/ is_sha3_slow.html. [Accessed 15 November 2018].

[14] P. Leach, M. Mealling and R. Salz, "A Universally Unique IDentifier (UUID) URN Namespace," Internet Engineering Task Force, July 2005. [Online]. Available: https://tools.ietf.org/html/rfc4122.html#section-4.1. [Accessed 15 November 2018].

[15] H. Delfs and H. Knebl, Introduction to Cryptography: Principles and Applications, Springer, 2015.

[16] A. Corbellini, "Elliptic Curve Cryptography: a gentle introduction," 17 May 2015. [Online]. Available: http://andrea.corbellini.name/2015 /05/17/elliptic-curve-cryptography-a-gentle-introduction/. [Accessed 15 November 2018].

[17] M. Bafandehkar, S. Yasin, R. Mahmod and Z. Hanapi, "Comparison of ECC and RSA algorithm in resource constrained devices," in International Conference on IT Convergence and Security, 2013.

Mobile Cloud Learning based on User Acceptance using DeLone and McLean Model for Higher Education

Kristiawan Nugroho[1], Sugeng Murdowo[2]
Computerized Accounting
AMIK Jakarta Teknologi Cipta
Semarang,Indonesia

Farid Ahmadi[3], Tri Suminar[4]
Faculty of Education
Universitas Negeri Semarang
Semarang,indonesia

***Abstract*—Mobile learning has been used in the learning process in several tertiary institutions in Indonesia. However, several universities have not been able to implement mobile learning due to the limitations of computer and network infrastructure. Cloud computing is a solution for agencies that experience limitations in computer infrastructure in the form of internet-based services for their customers. This paper discusses the implementation of mobile cloud learning which is a combination of mobile learning technology and cloud computing using the renewal DeLone and McLean Model which is a successful model to measure how important the implementation of a mobile cloud learning system is. The results showed that from the F test results obtained Fcount of 13,222, then Fcount> Ftable (13,222> 3.01), then Ho was rejected and H1 was accepted. So it can be concluded Information Quality, System Quality, and Service Quality together affect the Intensity of Use.**

Keywords*—*Mobile learning; cloud computing; DeLone; McLean; model

I. Introduction

The use of information technology in education has been widely carried out by various universities throughout the world. Elearning is an electronic learning tool as a learning aid tool that has been successfully applied in universities in several countries including Yemeni University in Yamen [1]. Today there has been a transition from technological forms from web base applications to mobile applications makes it easier for students to get learning wherever they are. Mlearning is a vital component of educational technology in higher education, students can learn, collaborate, share ideas about lessons using technology development on the internet [2]. Students can do learning activities without having to be in class, they can learn and interact with the material and lecturers whenever and wherever they are. The application of mobile learning technology is not only for ordinary students, but Mlearning is also applied to students who have dyslexic disorders. By adjusting multimodal functionality, Mlearning can improve the learning abilities of students with dyslexia by almost 30% in learning activities [3]. Tablets, PDAs and smartphones have become a part of human life, this equipment helps human activities, especially in the field of education. The use of mobile learning in modern learning is increasing, the increasing use of smartphones, the development of cellular networks and the global internet are increasingly increasing learning acceptance by using electronic means[4].

The use of mobile learning applications has been applied from the level of kindergarten education[5] up to undergraduate education[6]. Some countries such as South Korea, the US, Japan, Taiwan, Malaysia, Singapore, the European Union, and Australia have been able to utilize mobile devices in a learning environment quite effectively, they have benefited from increased efficiency to ease of accessibility of education to remote areas[7]. Research on Mobile Learning is growing by combining with Cloud Computing (CC) technology which is a distributed computer group (mostly data centers and servers) that provide demand resources and services using internet media[8], CC is a new technology paradigm that helps provide infrastructure with the specifications needed by an organization including universities.

The three services provided by CC are Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (Saas). By using the three CC infrastructures above, a university does not have to have infrastructure such as servers and storage areas. big data in running online learning. The development of research on Mobile Learning using Cloud Computing has been applied in various fields of life. Research on mobile cloud computing among others was conducted by Dhanalakshmi who discussed several advantages of using mobile cloud computing compared to traditional mobile learning[9]. In addition, research on the use of mobile cloud computing is also conducted on students who have dyslexia [3], They build interactive multimodal interfaces to facilitate dyslexia students in learning. In addition to the world of education, Mobile cloud learning is also applied in the world of health. Several types of research are to monitor the risk of coronary heart disease sufferers using the Adaptive Neuro Fuzzy Inference System / ANFIS method [10], MCC implementation is also carried out in stroke patients monitoring using classification techniques in Artificial Neural Network (ANN) which produces applications to monitor the condition of stroke sufferers with an Android-based application[11]. Problems and challenges in implementing mobile learning include the high costs associated with equipment, connectivity, maintenance, technical support [12].

This paper discusses the design and implementation of mobile learning using cloud computing technology known as Mobile Cloud Learning in universities in Indonesia using the Delone and Mclean (D&M) method. The D&M method is used to find out the most influential factors in the use of Mobile Cloud Learning (MCL) so that the implementation of MCL can be optimally implemented in learning in tertiary institutions.

II. Related Works

A. Mobile Learning

Mobile learning is a modern learning technique that can be done through wireless devices such as mobile phones, PDAs, laptops and hand held devices[13]. Various researches on mobile learning have been carried out using mobile devices including PDAs[14, p.], however limited screen and memory factors hamper enough comfort in the learning process. Research on the weather carried out by Cloughet al. shows that statistically the use of mobile learning has proven to be effective compared to studies that only use survey methods alone[15]. In other studies of mobile learning involving students Min Juan Wang stated that learning activities using mobile learning are better than conventional methods due to intellectual, emotional and behavioral involvement in using this technology learning[16].

The satisfaction of educated participants and their independence in using system functions in mobile learning have also been implemented in a learning activity involving 152 participants[17] In 2011 Gwo-Jen Hwang conducted an experiment on a course containing local culture in Taiwan regarding effectiveness in a learning environment that uses mobile learning. The results of the research show that the use of Mlearning can increase student interest and learning achievement [8]. Learning with mobile learning is also done by forming Whatsapp groups that for 40 days interact in the learning process[18]. The results of the study show that students according to students learning with Whatspp Mlearning are interesting and beneficial to them.

Research on the implementation of mobile learning conducted in Oman involving students from 56 universities also shows that the level of awareness and acceptance of students shows a good level which means that students are interested in using smartphones in learning[19]. Kattayat in 2017 conducted research on the implementation of mobile learning in a number of students. The results showed that learning techniques with mlearning could improve achievement[20].

B. Mobile Cloud Computing

Mobile Cloud Computing (MCC) is a technology concept that combines cloud computing technology, secular technology and computer network utilization [21]. Khan stated that since 2008 there were 42.8 million mobile cloud computing subscribers who are expected to grow to 998 million subscribers in 2014 [22]. Several studies that have been carried out in utilizing cloud computing technology include Kotwal stating that MCC is an amalgamation of mobile technology with cloud computing that provides cloud-based services for web and mobile users[23], Other studies conducted by Kumar and Rajalakshmi stated that MCC technology has focused on cloud security issues that exist in a variety of mobile devices[24]. Shamim believes that in a study around 2015 there would be 240 million businesses that use cloud computing services via mobile devices which would generate revenue in the MCL business of $ 5.2 billion[25]. The architecture of mobile cloud computing can be seen in Fig. 1.

C. DeLone and McLean Model (D&M Model)

The Delone and Mclean model is an information system model that was created in 1992[26]. D&M Model is a framework that functions to measure complicated dependent variables in research on information systems, in this method will be discussed about the influence of system quality and information quality on user usage and satisfaction[27]. Some research using the D&M model, among others, was carried out by Lee who examined what factors influenced the trust and satisfaction with the implementation of mobile banking in Korea[28]. This study uses the D&M model in assessing 3 factors: system quality, information quality and interface design quality, using 276 valid questionnaires from mobile Bangking customers. The results of the study show that system quality and information quality significantly influence customer trust and satisfaction while design quality the interface does not affect customer trust and satisfaction. In the field of health, Bossen examines the methodological approach and the results of the evaluation of electronic health records (EHR) at hospitals in Denmark[29]. The results of the study explain that overall the staff has a positive impression in implementing the EHR, Performance, reliability, time response, login and all support can be received. In 2017 Mahmoodi applied the D&M method in the field of education, the Virtual education system (VES) was evaluated using the D&M model, resulting in a conclusion that the quality of the system had a large impact. the level of 68.8% [30].

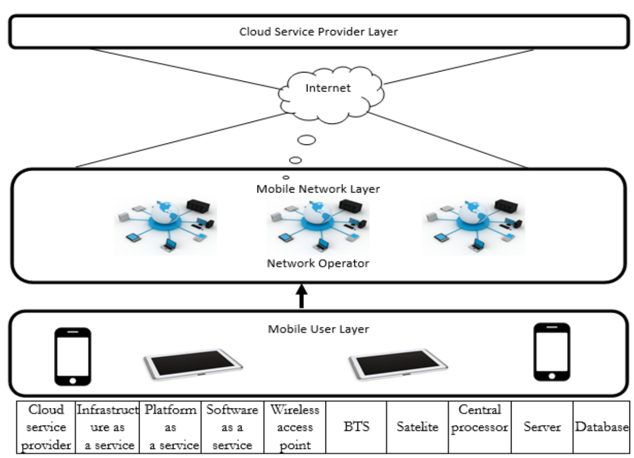


Fig. 1. The Mobile Cloud Computing Architecture (Source: Talal H. Noor, 2018).

III. System Design and Implementation

A. Proposed Method

The proposed method in the research regarding the implementation of mobile cloud learning can be seen in Fig. 2.

Fig. 2 above is the methodology used in research on the implementation of mobile cloud learning. Students and lecturers who are users of the system access the application using smartphones or tablets through the public network that can be accessed via data packages or wifi. Application is a combination of the concept of mobile learning in which there are several learning facilities in the form of downloading and uploading material, forums and quizzes combined with the concept of cloud computing which can use 3 services, namely, SaaS (Software as a Service), PaaS (Platform as Service) and IaaS (Infrastructure as a Service). This mobile application is evaluated using the DeLone and McLean Model to measure the level of acceptance of applications which include several factors namely Information Quality, System Quality, Intention to Users, User Satisfaction and Net Benefits.

B. System Design

The design of mobile cloud learning systems used OOA (Object Oriented Analysis) which is a modeling technique that combines data and processes in an object. The development of OOA techniques in this application uses Unified Modeling Language (UML) which is described in the form of use case diagrams, sequence diagrams and class diagrams

C. Design of System

The design of a mobile cloud learning application is illustrated in the use case diagram described in Fig. 3.

Fig. 3 above shows the relationship between three actors, namely student, lecture and admin who interact with each other in a system consisting of registration processes, learning materials, online studies, discussions, quizzes, examinations, smart notifications and teleconferences.

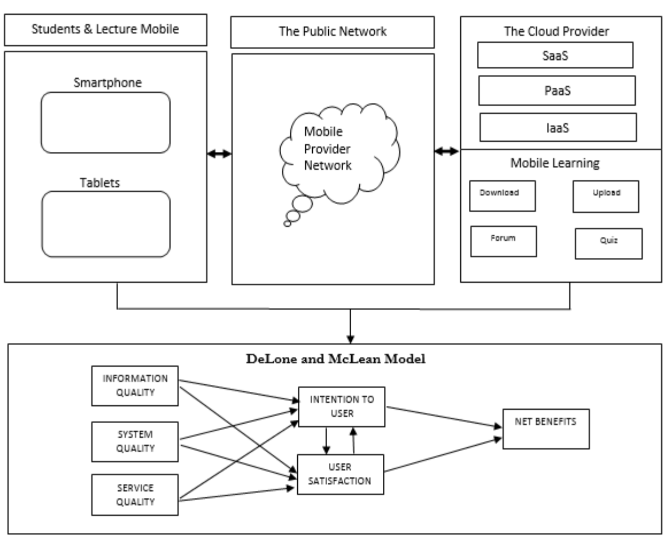


Fig. 2. Mobile Cloud Learning.

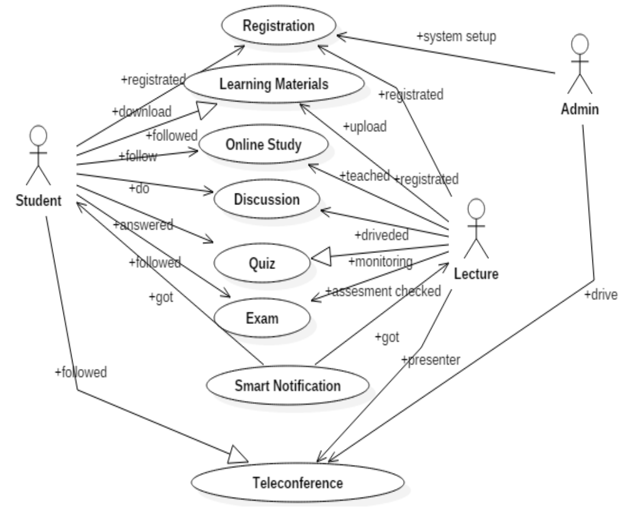


Fig. 3. Use Case Diagram.

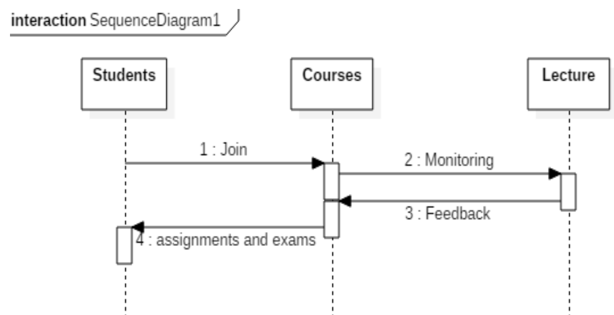


Fig. 4. Sequential Diagram.

Fig. 4 shows the design of a sequential diagram which includes the process beginning with students who join the courses held by each lecturer whose results will be evaluated with a quiz.

The class diagram of the mobile cloud learning system in Fig. 5 shows the relationship of classes that make up a system. Classes formed in this system include students, lectures, courses, discussions, online studies and quiz.

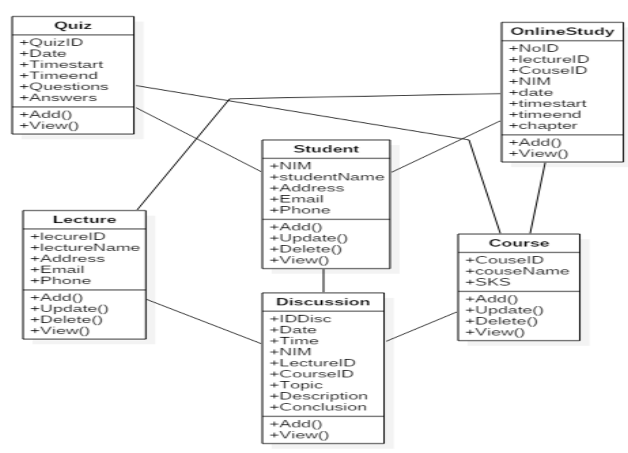


Fig. 5. Class Diagram.

IV. Result and Discussion

A. Data Collections

This study aims to determine the factors that influence the implementation of mobile cloud learning in universities in Indonesia. System testing is done by distributing questionnaires to the faculties of education outside the Universitas Negeri Semarang. Students run the mobile cloud learning application from the login process to the system dashboard, then students try to download material from the PNF Program Management course after which they work on a quiz consisting of several questions that must be answered one by one until the last question, the system will calculate the value achieved directly after students finish answering all questions and the results will be displayed in the system dashboard. After running the mobile cloud learning application students were distributed 28 questionnaires consisting of 32 questions to be answered where the questions contained regarding system quality, information quality, service quality, the intensity of use, user satisfaction and net benefits.

B. Program Implementation

The mobile cloud learning system is built using the Laravel framework and Android Studio web view. The results of making an application are shown in Fig. 6, 7, 8, 9 and 10 as follows.

Fig. 6 above shows the login application that must be filled in with a username and password that has been registered previously. If the user fills in a username and password that matches the one registered in the system, the user will be taken to the application's dahsboard menu display as shown in Fig. 7 as follows.

Mobile cloud learning users, namely students, can download material according to the class they are following as shown in Fig. 8.

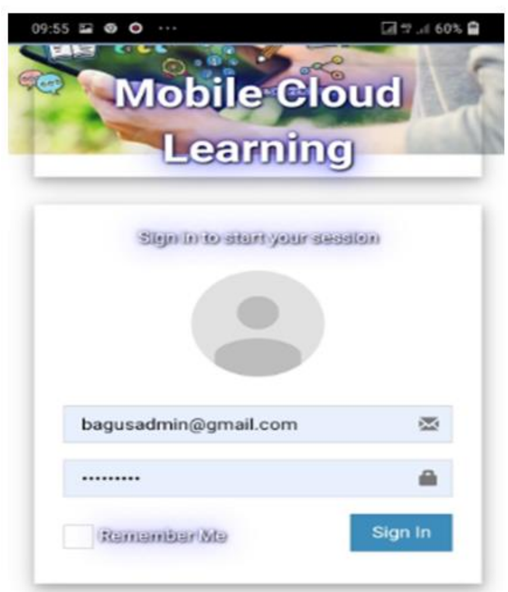


Fig. 6. Users Login.

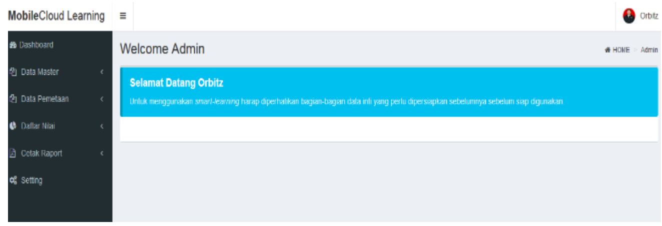


Fig. 7. Dashboard Menu.

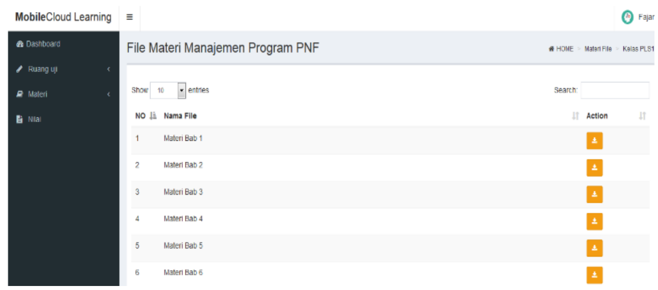


Fig. 8. Download Menu.

Fig. 9 shows a quiz application that can be answered by students after reading and studying the material they have downloaded previously. The questions in the quiz are randomized randomly according to the course taken by students.

After students answer all the questions in the quiz, the system will do the calculation and automatically give a value to the correct answer to the final result of the calculation as shown in Fig. 10.

C. System Testing

1) Test validity of measuring instruments: Validity shows the extent to which the measuring instrument is able to measure what you want to be measured. In testing the validity of using the Pearson Correlation technique, where the process is seen from the correlation value on the total value of the answers. The value in the answer row is compared with the r table with a significance value of 5% and the number of samples 28 rtable is 0.374 as shown in Table I.

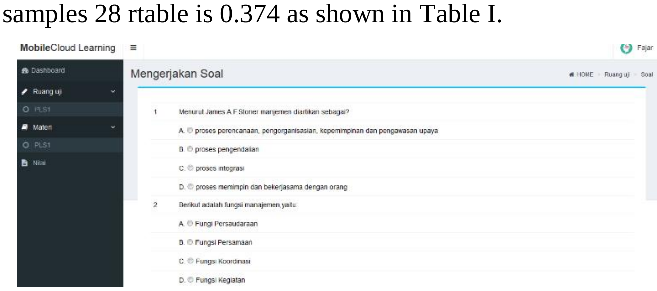


Fig. 9. Quiz Menu.

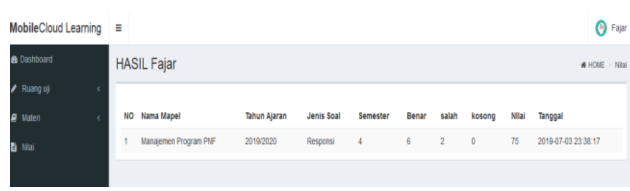


Fig. 10. Quiz Result.

TABLE. I. VALIDITY TEST RESULTS

Variable	Rcount	Rtable	Description
X11	0,762	0,374	Valid
X12	0,513	0,374	Valid
X13	0,716	0,374	Valid
X14	0,662	0,374	Valid
X15	0,751	0,374	Valid
X16	0,828	0,374	Valid
X21	0,774	0,374	Valid
X22	0,784	0,374	Valid
X23	0,81	0,374	Valid
X24	0,659	0,374	Valid
X25	0,305	0,374	Not Valid
X26	0,64	0,374	Valid
X27	0,578	0,374	Valid
X28	0,618	0,374	Valid
X31	0,66	0,374	Valid
X32	0,85	0,374	Valid
X33	0,835	0,374	Valid
X34	0,66	0,374	Valid
X35	0,865	0,374	Valid
X41	0,883	0,374	Valid
X42	0,76	0,374	Valid
X51	0,695	0,374	Valid
X52	0,837	0,374	Valid
X53	0,749	0,374	Valid
X54	0,56	0,374	Valid
X55	0,64	0,374	Valid
X61	0,589	0,374	Valid
X62	0,782	0,374	Valid
X63	0,584	0,374	Valid
X64	0,695	0,374	Valid
X65	0,525	0,374	Valid
X66	0,503	0,374	Valid

The test results are declared valid if rcount> from rtable, from the research variables there is 1 variable that is declared invalid, namely the Integrated Completeness Variable of the System Quality group

2) Measuring instrument reliability test: A reliability test in Table II is used to determine the consistency of measuring instruments that usually use a questionnaire. The reliability test is a continuation of the validity test where the items included in the test are valid items only. By using the 0.6 limits it can be seen whether the instrument is reliable or not. Provided less than 0.6 is not good, 0.7 is acceptable and above 0.8 is good.

TABLE. II. RELIABILITY TEST RESULTS

No	Variable	CronBach's Alpha	N of Items
1	Information Quality (X1)	0.881	6
2	System Quality (X2)	0.889	7
3	Service Quality (X3)	0.908	5
4	Intensity of Use (X4)	0.916	2
5	User Satisfaction (X5)	0.893	5
6	Net Benefits (Y)	0.846	6

3) Joint regression coefficient test (f test) 1: Joint regression tests are used to determine the effect together between the dependent variable (X) on the dependent variable (Y). The F Test results can be seen in Table III as follows:

a) Predictors (Independent): Information Quality, System Quality, Service Quality

b) Dependent Variabel: Intensity of Use.

Based on the F test results from the above table using a significance level of 0.05 and df 1 (number of variables -1) df1 = 4 - 1 = 3 and df2 (nk-1) or 28 - 3 - 1 = 24 (where n = number of respondents, k = number of independent variables) read from table F of 3.01.

For testing criteria if F arithmetic <F table, then Ho is accepted and if Fcount> Ftable then Ho is rejected.

From the F test results obtained Fcount of 13,222, then Fcount> Ftable (13,222> 3.01), then Ho is rejected and H1 is accepted. So it can be concluded Information Quality, System Quality, and Service Quality together affect the Intensity of Use.

4) Partial regression coefficient test (t test): T test is used to determine partially the independent variable (X) significantly influences the dependent variable (Y). Test results f using multiple linear regression can be seen in Table IV as follows:

a) Independent Variable

X1 – Information Quality

X2 – System Quality

X3 – Service Quality

b) Dependent variable Y (X4)– Intensity of Use

Test criteria T test is that if the tcount> ttable then H1 is accepted and Ho is rejected, and if the tcount <ttable then H1 is rejected and Ho is accepted.

TABLE. III. F TEST RESULTS

F Count Value
13.222

TABLE. IV. T TEST RESULT

Variable	t count	t table
X1	0.268	2.064
X2	2.040	2.064
X3	1.169	2.064

For the value of t table can be seen from the statistical table with a significance value of 0.05 / 2 = 0.025 and degrees of freedom df = 28-3-1 = 24 is 2.064.

Following are the results of the description of each hypothesis:

1) Information quality variable test results that: Information Quality does not affect the Intensity of Use obtained from the Mobile Cloud Learning System. This is evidenced by the value of tcount <ttable (0.268 <2.064) which means H1 is rejected.

2) System quality variable test results that system quality: does not affect the Intensity of Use obtained from the Mobile Cloud Learning System. This is evidenced by the value of tcount <ttable (2,040 <2,064) which means H1 is rejected.

3) Time service variable testing results that service: Quality does not affect the Intensity of Use obtained from the Mobile Cloud Learning System. This is evidenced by the value of tcount <ttable (1,169 <2,064) which means H1 is rejected.

5) The regression coefficient together test(test f) – 2: Joint regression tests are used to determine the effect together between the dependent variable (X) on the dependent variable (Y). The F Test results can be seen in table V as follows:

a) Predictors (Independent): Information Quality, System Quality, Service Quality, Intensity of Use

b) Dependent variabel: User Satisfaction

Based on the F test results from the above table using a significance level of 0.05 and df 1 (number of variables -1) df1 = 5-1 = 4 and df2 (nk-1) or 28 - 4 - 1 = 23 (where n = the number of respondents, k = number of independent variables) read from table F of 2.8. For testing criteria if F arithmetic <F table, then Ho is accepted and if Fcount> F table then Ho is rejected.

Based on the F test results obtained Fcount of 9,255, then Fcount> Ftable (9,255> 2.8), then Ho is rejected and H1 is accepted. So it can be concluded Information Quality, System Quality, Service Quality and Intensity of Use together affect the User Satisfaction.

6) Partial regression coefficient test (t test): T test is used to determine partially the independent variable (X) significantly influences the dependent variable (Y). Test results f using multiple linear regression can be seen in Table VI as follows:

a) Independent variable

- X1 - Information Quality
- X2 - System Quality
- X3 - Quality of Service
- X4 - Intensity of Use

b) Dependent Variable Y (X5)– User Satisfaction

Test criteria T test is that if the tcount> ttable then H1 is accepted and Ho is rejected, and if the tcount <ttable then H1 is rejected and Ho is accepted.

The value of the t table can be seen from the statistical table with a significance value of 0.05 / 2 = 0.025 and degrees of freedom df = 28-4-1 = 23 is 2.069.

Following are the results of the description of each hypothesis:

1) Information quality variable test results that: Information Quality does not affect User Satisfaction obtained from the Mobile Cloud Learning System. This is evidenced by the value of tcount <ttable (0.811 <2.069) which means H1 is rejected.

2) System quality variable test results that system quality: does not affect User Satisfaction obtained from the Mobile Cloud Learning System. This is evidenced by the value of tcount <ttable (-1.145<2.069) which means H1 is rejected.

3) Time service variable testing results that service: Quality does not affect User Satisfaction obtained from the Mobile Cloud Learning System. This is evidenced by the value of tcount <ttable (1,157<2,069) which means H1 is rejected.

4) Usage intensity variable testing results that usage: Intensity affects the User Satisfaction obtained from the Mobile Cloud Learning System. This is evidenced by the value of t> t table (3.094> 2.069) which means that Ho is rejected

7) Joint regression coefficient tes (test f) – 3: Joint regression tests are used to determine the effect together between the dependent variable (X) on the dependent variable (Y). The F Test results can be seen in Table VII as follows:

a) Predictors (Independent): Usage Intensity, User Satisfaction

b) Dependent Variable: Net Benefits

Based on the F test results from the Table VII using a significance level of 0.05 and df 1 (number of variables -1) df1 = 3 - 1 = 2 and df2 (nk-1) or 28 - 2 - 1 = 25 (where n = number of respondents, k = number of independent variables) read from table F of 3.39.

TABLE. V. F TEST RESULT

F Count Value
9.255

TABLE. VI. T TEST RESULT

Variable	t count	t table
X1	0.811	2.069
X2	-1.045	2.069
X3	1.157	2.069
X4	3.094	2.069

TABLE. VII. F TEST RESULT

F count value
18.137

For testing criteria if F arithmetic <F table, then Ho is accepted and if Fcount> F table then Ho is rejected.

From the F test results obtained Fcount of 18.137, then Fcount> Ftable (18.137> 2.8), then Ho is rejected and H1 is accepted. So it can be concluded that Intensity of Use and User Satisfaction jointly influences with Net Benefits.

8) Partial regression coefficient test (t test): T test is used to determine partially the independent variable (X) significantly influences the dependent variable (Y). Test results f using multiple linear regression can be seen in Table VIII as follows:

TABLE. VIII. T TEST RESULT

Variable	t count	t table
X4	0.682	2.060
X5	3.406	2.060

a) Independent Variable

X4 – Intensity of Use

X5 - User Satisfaction

b) Dependen Variable Y (X6)– Net Benefits

Test criteria T test is that if the tcount> ttable then H1 is accepted and Ho is rejected, and if the tcount <ttable then H1 is rejected and Ho is accepted.

For the value of t table can be seen from the statistical table with a significance value of 0.05 / 2 = 0.025 and degrees of freedom df = 28-2-1 = 25 is 2.0.

Following are the results of the description of each hypothesis:

1) User intensity variable test results that user intensity: does not affect the Net Benefits obtained from Sthe Mobile Cloud Learning System. This is evidenced by the value of t <t table (0.682 <2.060) which means H1 is rejected.

2) User satisfaction variable test results that user: Satisfaction affects the Net Benefits obtained from the Mobile Cloud Learning System. This is evidenced by the value of t> t table (3.406 <2.060) which means H1 is accepted.

V. CONCLUSION AND SUGGESTIONS

A. Conclusions

Based on the f and t test results it can be seen that Information Quality, System Quality, Service Quality, together affect the Intensity of Use in the Mobile Cloud Learning System process, while individually do not affect the Intensity of use in the Mobile Cloud Learning System implementation.

By used f and t test results it can be seen that Information Quality, System Quality, Service Quality, and Intensity of Use together affect user satisfaction. Information quality, system quality and service quality do not affect User in the Mobile Cloud Learning System implementation.

By calculated f and t test results it can be seen that Intensity of Use and user satisfaction together affect Net Benefits, user intensity does not affect Net Benefit. User Satisfaction influences Net Benefits in the Mobile Cloud Learning System implementation.

B. Suggestions

Mobile cloud learning needs to be improved, especially on smart notification and teleconference features. In addition, this system also needs to be hosted on a server that has large storage and bandwidth specifications so that multimedia files can run well.

ACKNOWLEDGMENTS

We would like thank to the Ministry of Research, Technology, and Higher Education for granted funding through Penelitian Kerjasama Perguruan Tinggi (PKPT), No 084/L6/AK/SP2H/PENELITIAN/2019.

REFERENCES

[1] A. S. A. Y. A.-A. Ivelina Peneva and Krasimir Yordzhev, “Student’s Readiness for E-learning in the Universities in Yemen,” Science and Technology Publishing (SCI & TECH), vol. 1, no. 8, August. 2017.

[2] M. Al-Emran, H. M. Elsherif, and K. Shaalan, “Investigating attitudes towards the use of mobile learning in higher education,” Computers in Human Behavior, vol. 56, pp. 93–102, Mar. 2016, doi: 10.1016/j.chb.2015.11.033.

[3] W. G. Alghabban, R. M. Salama, and A. H. Altalhi, “Mobile cloud computing: An effective multimodal interface tool for students with dyslexia,” Computers in Human Behavior, vol. 75, pp. 160–166, Oct. 2017, doi: 10.1016/j.chb.2017.05.014.

[4] A. Chavoshi and H. Hamidi, “Social, individual, technological and pedagogical factors influencing mobile learning acceptance in higher education: A case from Iran,” Telematics and Informatics, vol. 38, pp. 133–165, May 2019, doi: 10.1016/j.tele.2018.09.007.

[5] A. Drigas, G. Kokkalia, and A. Economou, “Mobile Learning For Preschool Education,” International Journal of Interactive Mobile Technologies (iJIM), vol. 10, no. 4, p. 57, Oct. 2016, doi: 10.3991/ijim.v10i4.6021.

[6] B. Ferdousi and J. Bari, “Infusing Mobile Technology into Undergraduate Courses for Effective Learning,” Procedia - Social and Behavioral Sciences, vol. 176, pp. 307–311, Feb. 2015, doi: 10.1016/j.sbspro.2015.01.476.

[7] A. I. Khan, H. Al-Shihi, Z. A. Al-khanjari, and M. Sarrab, “Mobile Learning (M-Learning) adoption in the Middle East: Lessons learned from the educationally advanced countries,” Telematics and Informatics, vol. 32, no. 4, pp. 909–920, Nov. 2015, doi: 10.1016/j.tele.2015.04.005.

[8] G.-J. Hwang and C.-C. Tsai, “Research trends in mobile and ubiquitous learning: a review of publications in selected journals from 2001 to 2010: Colloquium,” British Journal of Educational Technology, vol. 42, no. 4, pp. E65–E70, Jul. 2011, doi: 10.1111/j.1467-8535.2011.01183.x.

[9] D. S. Dhanalakshmi, S. Suganya, and K. Kokilavani, “Mobile Learning Using Cloud Computing,” vol. 2, p. 7, 2014.

[10] C. Venkatesan, P. Karthigaikumar, and S. Satheeskumaran, “Mobile cloud computing for ECG telemonitoring and real-time coronary heart disease risk detection,” Biomedical Signal Processing and Control, vol. 44, pp. 138–145, Jul. 2018, doi: 10.1016/j.bspc.2018.04.013.

[11] Y. Karaca, M. Moonis, Y.-D. Zhang, and C. Gezgez, “Mobile cloud computing based stroke healthcare system,” International Journal of Information Management, vol. 45, pp. 250–261, Apr. 2019, doi: 10.1016/j.ijinfomgt.2018.09.012.

[12] Y. Mehdipour and H. Zerehkafi, “Mobile Learning for Education: Benefits and Challenges,” p. 9.

[13] S. AhmadBhat and S. Al Saleh, “Mobile Learning: A Systematic Review,” IJCA, vol. 114, no. 11, pp. 1–5, Mar. 2015, doi: 10.5120/20019-1406.

[14] J. Schreurs, “Mobile e-learning course scenario model on PDA,” IJMLO, vol. 2, no. 4, p. 358, 2008, doi: 10.1504/IJMLO.2008.020688.

[15] G. Clough, A. C. Jones, P. McAndrew, and E. Scanlon, "Informal learning with PDAs and smartphones: Informal learning with PDAs," Journal of Computer Assisted Learning, vol. 24, no. 5, pp. 359–371, Sep. 2008, doi: 10.1111/j.1365-2729.2007.00268.x.

[16] M. Wang, R. Shen, D. Novak, and X. Pan, "The impact of mobile learning on students' learning behaviours and performance: Report from a large blended classroom," British Journal of Educational Technology, vol. 40, no. 4, pp. 673–695, Jul. 2009, doi: 10.1111/j.1467-8535.2008.00846.x.

[17] S.-S. Liaw, M. Hatala, and H.-M. Huang, "Investigating acceptance toward mobile learning to assist individual knowledge management: Based on activity theory approach," Computers & Education, vol. 54, no. 2, pp. 446–454, Feb. 2010, doi: 10.1016/j.compedu.2009.08.029.

[18] T. Bansal and D. D. Joshi, "A Study of Students' Experiences of Mobile Learning," p. 8, 2014.

[19] M. Sarrab, "M-learning in Education: Omani Undergraduate Students Perspective," Procedia - Social and Behavioral Sciences, vol. 176, pp. 834–839, Feb. 2015, doi: 10.1016/j.sbspro.2015.01.547.

[20] S. Kattayat, S. Josey, and A. J.V, "Mobile Learning Apps in Instruction And Students Achievement," Int. J. Interact. Mob. Technol., vol. 11, no. 1, p. 143, Jan. 2017, doi: 10.3991/ijim.v11i1.6420.

[21] E. Ibukun and O. Daramola, "A Systematic Literature Review of Mobile Cloud Computing," IJMUE, vol. 10, no. 12, pp. 135–152, Dec. 2015, doi: 10.14257/ijmue.2015.10.12.15.

[22] A. N. Khan, M. L. Mat Kiah, S. U. Khan, and S. A. Madani, "Towards secure mobile cloud computing: A survey," Future Generation Computer Systems, vol. 29, no. 5, pp. 1278–1299, Jul. 2013, doi: 10.1016/j.future.2012.08.003.

[23] Priya. A. Kotwal and Adwitiy. R. Singh, "Evolution and effects of mobile cloud computing, middleware services on cloud, future prospects: A peek into the mobile cloud operating systems," in 2012 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, India, 2012, pp. 1–5, doi: 10.1109/ICCIC.2012.6510270.

[24] R. Kumar and S. Rajalakshmi, "Mobile Cloud Computing: Standard Approach to Protecting and Securing of Mobile Cloud Ecosystems," in 2013 International Conference on Computer Sciences and Applications, Wuhan, China, 2013, pp. 663–669, doi: 10.1109/CSA.2013.161.

[25] S. MShamim, A. Sarker, A. Newaz Bahar, and Md. Atiqur Rahman, "A Review on Mobile Cloud Computing," IJCA, vol. 113, no. 16, pp. 4–9, Mar. 2015, doi: 10.5120/19908-1883.

[26] W. H. DeLone and E. R. McLean, "Information Systems Success: The Quest for the Dependent Variable," Information Systems Research, vol. 3, no. 1, pp. 60–95, Mar. 1992, doi: 10.1287/isre.3.1.60.

[27] H. Wang and Z. Hu, "Applying DeLone and McLean Information System Model to Online Consumer Behavior in China," in 2009 International Conference on Management of e-Commerce and e-Government, Nanchang, China, 2009, pp. 74–77, doi: 10.1109/ICMeCG.2009.86.

[28] K. C. Lee and N. Chung, "Understanding factors affecting trust in and satisfaction with mobile banking in Korea: A modified DeLone and McLean's model perspective," Interacting with Computers, vol. 21, no. 5–6, pp. 385–392, Dec. 2009, doi: 10.1016/j.intcom.2009.06.004.

[29] C. Bossen, L. G. Jensen, and F. W. Udsen, "Evaluation of a comprehensive EHR based on the DeLone and McLean model for IS success: Approach, results, and success factors," International Journal of Medical Informatics, vol. 82, no. 10, pp. 940–953, Oct. 2013, doi: 10.1016/j.ijmedinf.2013.05.010.

[30] H. Irawan and I. Syah, "Evaluation of implementation of enterprise resource planning information system with DeLone and McLean model approach," in 2017 5th International Conference on Information and Communication Technology (ICoIC7), Melaka, Malaysia, 2017, pp. 1–7, doi: 10.1109/ICoICT.2017.8074721.

Open Challenges for Crowd Density Estimation

Shaya A. Alshaya
Computer Science department
College of Sciences and Humanities at alGhat
Majmaah University, Saudi Arabia

***Abstract*—Nowadays, many emergency systems and surveillance systems are related to the management of the crowd. The supervision of a crowded area presents a great challenge especially when the size of the crowd is unknown. This issue presents a point of start to the field of the estimation of the crowd based on density or counts. The density of a crowded area is one of the important topics dealt with in many kinds of applications like surveillance, security, biology, traffic. In this paper, we try not only to present a deep review of the different approaches/techniques used in the previous works to estimate the size of the crowd but also to describe the different datasets used. A comparison of some related works based on the weakness and the strength features of each approach is highlighted to show the important research key related to the field of the estimation of the crowded area.**

***Keywords*—Crowd density; count density; deep learning; CNN; datasets; metrics**

I. Introduction

The surveillance system is widely used in our life daily. It is mounted in bank agency [1], traffic [2], mall [3], etc. Their uses are different from one application to another. This paper is focused in the use of the surveillance systems to estimate the size of the crowd.

Estimate crowd density aims to understand the behavior of a crowded scene and to analyze the comportment for better security, management, and safety. A computer vision technique is used for all systems. Some of them propose the computation on a single image, and the most compute frames from a video streaming. Crowd analysis is associated with multiple disciplinary research topic as computer vision, public safety, biology [4], physics [5], psychology [6].

Crowd density estimation can be classified into five research topics.

1) Disaster management: Systems based on the estimation of the crowd aims to supervise the behavior to avoid disasters in several situations as music concerts, sports events, political rallies, and public demonstrations.

2) Safety monitoring: Crowd analysis help in understanding behavior, congestion, anomaly, and event [7]. These analyses are applied for many video surveillance purposes such as shopping malls, airports, and sports events.

3) Design of public spaces: Crowd analysis improves the optimization of public spaces design to ensure more safety in crowded situations.

4) Intelligence gathering and analysis: Crowd analysis is used for interesting products, interesting places. It ensures intelligence in queuing systems. Therefore, analysis improves the knowledge of the system and helps in improvement or optimization strategies.

5) Forensic search: Crowd analysis determine a particular data in a crowded scene as detecting suspicious behavior or detecting suspects [8], [7].

These topics have encouraged researchers in different specialties to contribute and to improve the estimation of a crowded area via various methods and related tasks such as density estimation [9] counting[9], [10, 11], tracking [11], behavior analysis [12]. All these tasks can be extracted from a crowded scene and there can be applied for different applications. The challenge is increased when the scene is identified as a very high dense situation. Previous studies use a variety of techniques/methods like regression [13], clustering [14], and detection [15] to count or to estimate crowds. These approaches require standards dataset to estimate the performance of crowd density analysis.

This paper is distributed as follows: different datasets used by researchers to evaluate their approaches are described in Section 2. A review of crowd density estimation approaches is presented in Section 3. A comparison between previous approaches in crowd density estimation is performed in Section 4. Finally, a conclusion based on open challenges is presented in the last section.

II. Datasets

In vision processing systems, datasets represent an essential requirement to evaluate their proposed design. This section lists the different datasets used by the previous works to assess the estimation of the crowded area approaches. Some related works perform their own dataset but most of the studies use standard and universal datasets. We focus on this section to the standard dataset used in the field of the crowded zone.

- WorldExpo’10 dataset [16]: This dataset is characterized by their size. It is composed of a big number of scene performed for the count of the crowd. It is characterized by the number of prototypes (1132 video clips), the number of scenes (108), the resolution (576*720) which are bigger than other datasets. These videos are captured by more than one hundred cameras.

- UCSD dataset [17]: It includes 2000 frames extracted from video streaming. This dataset is limited in terms of scenes because all frames are done by one camera which means one scene. Frames have little resolution (158*238).

TABLE. I. CHARACTERISTICS OF EACH DATASET

Datasets		Number of frames	Resolution	Number of scenes	Number of frames/second	Minimum number of persons	Maximum number of persons	Total number of persons	Average crowd count
WorldExpo'10		4440000	576*720	108	50	1	253	199923	50.2
UCF_CC_50		50	-	50	1	94	4543	63974	1279.5
UCSD		2000	158*238	1	10	11	46	49885	24.9
ShanghaiTech	**Part A**	482	different	-	-	33	3139	241677	501.4
	Part B	716	768*1024	-	-	9	578	88488	123.6

- UCF_CC_50 [18]: This dataset includes a limited number of frames (50) but the number of labeled pedestrians is bigger than the UCSD dataset and it achieved about 63000.
- ShanghaiTech dataset [19]: this dataset introduces a large scale crowd. It includes 330165 people as the total number of labeled pedestrians. It includes 1198 images. These images define two groups: The Part A is grouped randomly from images stored in the Internet, and the part B is composed of images captured from Shanghai streets.
- Make3D [18]: the resolution of this dataset is 2272*1704. This dataset is adopted to learn features and it estimates the scene depth from a single frame. The Make3D dataset provides more than 1000 scenes composed of outdoor and indoor scenes.

In light of this brief review, we mention that each dataset could be applied for a particular case. Table I resumes characteristics of the most used datasets in literature.

III. EVALUATION METRICS

This section discusses the different metrics used to evaluate crowded systems. In the literature, there are four factors. The most two factors used by them are:

- The Mean Absolute Error (MAE) [20] computes the average of all absolute errors which is defined by theses following formula:

Absolute Error (AE) computes the error rate between the true value (x) and the measured value (xi) related to n frames.

$$AE = |X_i - X| \quad (1)$$

The MAE is the average of all absolute errors:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - X| \quad (2)$$

- The Mean Squared Error (MSE) [20] represents the average of all errors related to the distance between the regression line and the value. The regression line is the best line drawn by the measured data. The accuracy is higher when the MSE is smaller.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - X)^2 \quad (3)$$

Other studies use the following metrics to evaluate crowded systems:

- Mean Windowed Relative Absolute Errors MWRAE [21] computes the average of all errors related to the distance between the real counts and the estimated counts. This metric is defined by the following formula:

$$MWRAE = \frac{1}{n}\sum_{i=1}^{n}\frac{\|C_i - \widetilde{C_i}\|}{C_i} \times 100 \quad (4)$$

Where Ci is the real count of the crowd related to the ith test video stream, $\widetilde{C_i}$ is the estimated count of the crowd related to the ith test video stream, and the parameter n represents the total number of test stream.

- Root MSE (RMSE) [17] computes the averages of all errors related to the standard deviation. This metric defines the best line around data based on the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - X)^2} \quad (5)$$

Where the xi is the measured value and the x is the true value of n frames.

IV. REVIEW OF CROWD DENSITY ESTIMATION METHODS

This study is based on a deep search on Web of science database since December 2017. The most significant keywords for this search are 'Crowd density estimation' that describes the scope of this paper. During the study collection, we set only papers written in English and dealt about the density/count estimation a crowd.

During the search, we use the combination of the following words: "Crowd", Density Estimation", Crowd Count" to find papers related to the scope. Logical operators are used between keywords. Only articles on journals and conferences are approved. Steps of the selection of the paper can be resumed as follow:

First step: This step aims to filter papers according to the title and the abstract.

Second step: Eliminate duplicate papers that use the same methods in the same dataset.

Third step: In this step, we approve papers agreed with necessary criteria: articles in English, known authors, reviewed paper, developing paper, and discussed paper.

In the literature, the crowded scene is seen either by the level of density or by counting the number of people.

S. Lin et al., [22], propose an intelligent algorithm based on SVM classifier to detect heads. The frame is proceeded by a processing phase to reduce noises. Then an extraction phase is performed by Haar wavelet transformation and normalization step. Finally, the matching phase ensured by the SVM classifier is chosen. The SVM aims to classify the extracted features belong to head class or not. The estimation accuracy is between 90-95% (about 125 persons in image). The experimentation shows that the camera position has to aligned to the optimal value of the angle (72.5 degrees). This angle is defined by the camera sensor position and the plane of the crowd. The method proposed by the authors supposes a unique size of all human heads and a uniform repartition of the crowd over the horizontal plane.

JH. Yin et al., [23], performs five methods to recognize the size of the crowd area. The first method removes the background based on the subtraction of the reference image and computes the occupied surface. The error associated with this method is about 15%. The second method computes the total perimeter of the busy area by applying the Edge detection algorithm. When the number of people has increased the accuracy of the algorithm is decreased. The error is about 23%. The third method combines the two-last method to improve estimation accuracy. The crowd density estimation error is decreased to 8%. These three methods still suffer from near-far effect especially. Persons how are near the camera occupied more area than other distant persons. Then authors propose the fourth method based on Geometric distortion to compensate for the near-far effect. The fifth method attempts to detect the movement without identifying objects in video streaming. This is done by the optical flow that is defined by the difference of the brightness from one image to the next. Based on Horn's optical flow algorithm, the motion is measured. These methods did not take into consideration the constraints of real-time execution.

CS. Regazzoni et al., [24], propose an estimation approach. They use temporal information of a sequence image. A means of a distributed Kalman filter network is performed. The proposed approach attempts to synchronize between multiple sensors based on modularity and data-fusion. The distributed Extended Kalman Filtering (DEKLF) algorithm implements both static models and status history. The first one is defined by some features of the edge function as the number of vertical edges, the number of edge points, the sum of the amplitudes of the maxima detected in the shape. The second one is defined by the depletion, the enhancement, and the steady conditions of the number of people. Algorithms are chosen to increase density accuracy and real-time exigence. The experimentation discusses the results according to a comparison between the proposed DEKF and the Bayesian belief network. The error is less than 20%.

AN. Marana et al., [25], use the Minkowski fractal dimension to estimate the count of people. This method verifies the case of a railway station. The edge detection is performed to the input image. Then a binarization step followed by a dilation method (enlarge the boundaries of regions) is applied to the image. Finally, the fractal dimension classifies the image according to the density into very high, high, moderate, low, and very low rubrics. The authors evaluate their method by comparing results with Minkowski methods and the Gray Level Dependence Matrix (GLDM) [26]. The last method is not able to distinguish between area with very high density and area with high density.

SY. Cho et al., [27] choose to apply the neural network as an intelligent algorithm to find an accurate result of the crowd's size. This method is based on background removal. The neural network is applied to identify if a mask belongs to black features, white features, and edge features. The case study is implemented for Railway station. This paper proposes a novel block diagram: a fast edge detection is proposed to ensure real-time. A binary step is applied instead of the Sobel filter. Then the edge algorithm is used. Then an estimation of the undesired region is performed by a crowd object extraction. This is done by removing the background. Finally, a Hybrid Global Learning (HGL) associated to a neural model is implemented. The HGL is performed by three algorithms. The first algorithm performs a hybrid of least squares and random search. Results prove that is the fast one with 2.02 min (CPU running time for learning) but the lowest estimation accuracy (90.72%). The second algorithm performs a hybrid of least squares and Simulated Annealing (SA). It obtains the best estimation accuracy (94.36%) but the worst speed (197.5 min). The third algorithm performs a hybrid of least squares and genetic algorithm (GA). It obtains 75.3 min in terms of time learning and 93.89 % for estimation accuracy.

C. Wang et al., [28] apply an end-to-end deep CNN regression to approximate people's number in a condensed crowd. The authors focus to decrease the influence of the ground by including negative samples to the training data. The truth counting of these samples is defined as zero. The proposed method enhances the estimation of counting persons. Results highlight a decrease of the error between absolute difference and the normalized absolute difference by 16.7% and 27% to mean respectively. The error rate is about 10 % which is still important. Nevertheless, this method is limited to 1300 persons per image.

F. Min, [29] presents an optimized method of the CNN named ConvNet to enhance the accuracy and the speed of the estimated crowd density. The author implements two stages on the cascade of CNN and he proposes to remove some network connections of the CNN design to speed up the computation. Experimentation is based on the PETS_2009 dataset. This method is limited to an image size of 42x40. Results show a decrease in the error rate to 3.2 %. Unfortunately, the author does not discuss the acceleration achieved by his method.

C. Zhang et al., [16] approximate the crowd's density/count especially for unseen scene by applying a deep CNN. The authors describe a data-driven method to finetune the trained CNN model for the target scene. They, also, built a novel dataset constituted with 108 frames which supports about 200,000 persons. A comparative study based on other datasets is done to show the reliability and the effectiveness of their method.

E. Wolf et al., [18] attempt to count persons by employing CNN. The addressed method focuses on layered boosting and selective samples. It aims to enhance accuracy and speed up the processing time. The authors achieve their goal by reducing the mean absolute error from 20% to 35% and the training time is decreased by 50%.

Z. Zhao et al., [21] present a CNN- based method to compute the number of persons across a line-of-interest. The method uses pairs of videos as inputs and it performs the training with pixel-level supervision maps. The proposed enhancement let the CNN learn more about features by decomposing the training phase into two steps: (1) Estimate the crowd density map, and (2) Estimate crowd velocity map. This decomposition provides more accuracy to solve the original problem by starting to answer each step. The authors perform a new dataset based on pedestrian trajectory annotations to show the robustness of the method via introducing a novel metrics: Mean Windowed Absolute error (6%).

Y. Zhang et al., [19] try to estimate the crowd from an unique image by performing a Multi-column CNN architecture. The MCNN supports any size or resolution of the input image. The method uses filters with different sizes to let CNN learn the features of each column. Then a geometry-adaptive kernel is used to compute the true density map associated with the input image. A new dataset including 1198 images is introduced by authors to cover all the challenging situations. Experimentation shows that the mean absolute error is 1.07%.

C. Shang et al., [30] attempt to count the crowd directly from an input image using an end-to-end CNN. The method estimates the crowd based on both global and local features by applying a pre-trained CNN to the image. The recurrent network layers provide the local counting by mapping features. The local count reduces the training time, and the global count enhances the accuracy of the results obtained by the local regions. A comparative study based on many databases is discussed to demonstrate the effectiveness of their attempt.

T. Mundhenk et al., [31] apply the deep learning method to count the crowd related to cars. The authors perform a large contextual dataset to help drivers to choose the best target and avoid bottlenecks. The proposed method aggregates residual learning and inception-style layers. This solution represents a new way to counts objects instead of the base of the known method on density estimation and localization. The authors prove via their experimentation that results are more accurate and the processing time is faster.

L. Boominathan et al., [32] announce the “crowdnet” framework based on the deep CNN to count the density of the crowd. Crowdnet is performed by the combination of the deep and shallow applied to a static image. This aggregation provides effective results associated with semantic information and features. To improve accuracy, the authors propose to enlarge the trained dataset to exceed 100 samples. Results are discussed using UCF CC 50 dataset.

A. Vishwanath et al., [33] count the crowd by using both the end-to-end cascaded CNN and the density map estimation. The proposed idea by authors provides the estimation of the crowd by classifying count into groups. This method enables us to learn globally features that refine highly the density maps and decreases the error count. A comparative study is highlighted to prove the accuracy of the density maps with the minimum count error.

S. Deepak et al., [34] present a mapping method between crowd counting and their density. A multi-scale CNN is described to decrease the worst effects of some factors as inter-occlusion, the high similarity of appearance, and view-points. The method is based on the switching of the CNN according to independent regressors to enhance the accuracy and the estimation. The proposed switch between classifiers to select the best CNN regressor. Results show that the switch relays patch to, particularly column in CNN to identify the crowd density of the input image. The comparative study proves that the proposed method enhances the accuracy and the mean error is decreased to near 2%.

X. Yang et al., [35] present an emergency evacuation as a case study related to the crowd area. The authors perform a clustering algorithm to extract informed and uninformed walkers. The goal of their study is to find the optimized guide during evacuation. The density of the crowd constitutes important criteria to achieve their goal. The informed method with an exponent model attains an approved accuracy.

Z. Zhikang et al., [9] propose to count the crowd based on many structures. The authors announced their method named the Adaptative Capacity Multi-scale CNN. This method ensures the assignment of different capacities to different portions. This method focuses on important regions instead of the whole image to ensure optimized allocations. The proposed method is composed of a fine network, a coarse network, and a smooth network. The first one finds the region to be focused and produce the rough feature map. The second one extracts the region of interest into a fine feature map. The third one enhances results by aggregate the two studied features to decrease the effect of division. The proposed method is well validated according to five used datasets.

Z.Liping et al., [36] introduce a deep learning technique to compute the crowd’s density in the case of non-uniform density and variations. The authors apply pooling operation to the density map to overcome the loss of the local spatial information. This pooling is performed by the use of dilated CNN to support details related to person position. This last feature is provided by global context guidance. The proposed method is proved by the use of many datasets.

X. Zeng et al., [37] attempt to decrease the problem of the scale variation related to the crowd’s estimation. The authors propose to provide more accurate contextual information by using a deep scale purifier network. The described method encodes multiscale features. The proposed supports a frontend and a backend model. A cross scene evaluation is applied to the approach. Many datasets are used to evaluate the accuracy of the DSPNet method.

This brief review proves that most techniques are applied only for an image. The authors in [18], [21], [33], and [33] propose an attempt to treat video instead the image to estimate the density of the crowd. These attempts should be enhanced to support any inputs. Recent works adopt deep learning methods to compute the density. These attempts request a learning phase based on a dataset. The high accuracy is the strong point of these methods but they suffer always from the increased time of processing. Datasets aim to evaluate the performance of methods proposed by researchers to estimate the crowd's size. When the evaluation is made by different datasets, results are more acceptable.

The real-time constraints are not well studied by the cited related works. The authors in [18], [21] attempt to propose methods with respect to the real-time constraints. This field requests to propose many hardware architectures to be implemented into a camera to estimate the density of the crowd.

This section has discussed some important studies related to the estimation of the density or the counts of crowds. The presented review lists many techniques based on video or image processing. Some methods extract the density according to the spatial information of the frame. These methods are accurate only in the case of the small size of crowds (inferior to 50). Other methods based on deep learning techniques show more accurate results especially in the case of the biggest size of crowds.

V. Synthesis

This section discusses the most important studies to extract the benefits and limits of each work. Then, a comparison based on different results metrics is highlighted to show the accuracy. At the end of this section, the evolution of this field is presented according to the number of publications during the last five years.

Previous works should be presented according to their characteristics, strengths, and weaknesses. Table II describes many studies based on the nature of the input, application type, used approach, used dataset, benefits, limits, and real-time processing.

TABLE. II. Overview of the Principal Methods to Estimate/Count Crowd Area

Authors	Inputs	Application	Approach	Dataset	Benefits	Limits	Real-Time
[28]	Image/different position	Pedestrians	End-to-end deep model CNN	UCFCC	Reduce the mean and the deviation of AD and NAD	Only 1300 per/image	NA
[29]	Image 42x40	Pedestrians	Optimized CNN (2 layers)	PETS-2009	Speed up processing and increase the correct rate	Complicate algorithm and limited dataset	NA
[16]	Image 158x238	Pedestrians	CNN	New dataset WorldExpo'10 UCSD UCF_CC_50	Applied for unseen scene	NA	NA
[18]	Image/Video 158x238	Bacterial cells, microscopy images	Gradient boosting with CNN (2 layers)	EXPO UCSD UCF50	Increasing accuracy	NA	Reduction by 50% in training time
[21]	Video 1280x720	Alley, street square	CNN with NVidia Titan GPU	New dataset	Decrease the time processing	A non-standard dataset is used	Yes, T=0.1s
[19]	Image/arbitrary camera	Pedestrians	Multi-column CNN	New dataset UCSD UCF_CC_50 WorldExpo'10	Improve estimation density	NA	NA
[30]	Image 640x480	Pedestrians	CNN based on local and global mapping	New dataset UCF_CC_50 WorldExpo'10	Decrease the time processing	NA	NA
[31]	Image 256x256	Cars	CNN with residual learning	New dataset	Increase accuracy	Time processing increased	
[32]	Image 225x225	Pedestrians	Deep learning and shallow	UCF_CC_50	Increase Accuracy	Insufficient number of training image	Yes
[33]	Image/Video	Pedestrians	End-to-end cascade CNN	ShanghaiTech UCF_Crowd_50	Increase Accuracy	NA	NA
[34]	Image/Video	Pedestrians	Switch-CNN	WorldExpo'10 UCSD ShanghaiTech UCF_CC_50	Increase Accuracy	NA	NA
[9]	Image	Pedestrians	ACM-CNN	WorldExpo'10 UCSD ShanghaiTech UCF_CC_50	Focus on important regions only	NA	NA
[36]	Image	Pedestrians	Dilated Convolution with Global Self-Attention	ShanghaiTech UCF_CC_50 UCSD	Decrease the loss of the low-level features	NA	NA
[37]	Image	Pedestrians	Deep scale purifier network	ShanghaiTech UCF_CC_50 UCF-QNRF	Reduce the loss of the contextual information	NA	NA

The following Tables III-VII compare the results of the related works according to the used datasets and the used metrics.

TABLE. III. COMPARATIVE STUDIES BASED ON FOUND RESULTS RELATED TO WORLDEXPO'10 CROWD COUNTING DATASET

Methods	Average MAE	Average MSE
Z. Zhikang et al., [9]	8.56	-
LBP+ RR [16]	31	17.4
Crowd CNN [16]	14.1	15.5
Fine-tuned Crowd CNN [16]	12.9	14.9
Crowd CNN+RR [16]	10.7	14.3
Sam et al., [34]	9.4	-
Regression learning [38]	26.7	14.3
Ridge Regression [39]	16.5	14.1
Shang et al., [40]	11.7	-
Y. Zhang et al., [41]	11.6	-

TABLE. IV. COMPARATIVE STUDIES BASED ON FOUND RESULTS RELATED TO UCSD DATASET

Methods	MAE	MSE	MWRAE
Z. Zhikang et al., [9]	**1.01**	**1.29**	-
C. Zhang et al., [16]	1.6	3.31	-
Walach et al., [17]	1.53	-	-
Zhuoyi et al., [21]	0.9	-	0.54
Sam et al., [34]	1.62	2.1	-
Z.Liping et al., [36]	1.08	1.44	-
Ridge Regression [39]	2.25	7.82	-
Y. Zhang et al., [41]	1.07	1.35	-
An et al., [42]	2.16	7.45	-
A. B. Chan et al., [43]	2.24	7.97	-
K. Chen et al., [44]	2.07	6.86	-
Lempitsky et al., [45]	1.7	-	-
Zhang et al., [46]	1.6	-	-
Pham et al., [47]	1.6	-	-
Ma et al., [48]	0.64	-	0.61

TABLE. V. COMPARATIVE STUDIES BASED ON FOUND RESULTS RELATED TO UCF_CC_50 DATASET

Methods	Mean Absolute Error	Mean Squared Error
Z. Zhikang et al., [9]	291.6	337
C. Zhang et al., [16]	467.0	498.5
Crowd CNN+RR [16]	467	498.5
Walach et al., [17]	474.0	-
Sam et al., [34]	318.1	439.2
Z.Liping et al., [36]	257.0	343.9
X. Zeng et al., [37]	243.3	307.6
Shang et al., [40]	270.3	-
Y. Zhang et al., [41]	377.6	509.1
Rodriguez et al., [49]	655.7	697.8
Learning to count [45]	493.4	487.1
Lempitsky et al., [45]	493.4	487.1
Zhang et al., [46]	467.0	-
MS counting [50]	468	590.3
Idrees et al., [50]	419.5	541.6
Boominathan et al., [51]	452.5	-
Sindagi et al., [52]	322.8	397.9
Onoro-Rubio et al., [53]	465.7	371.8

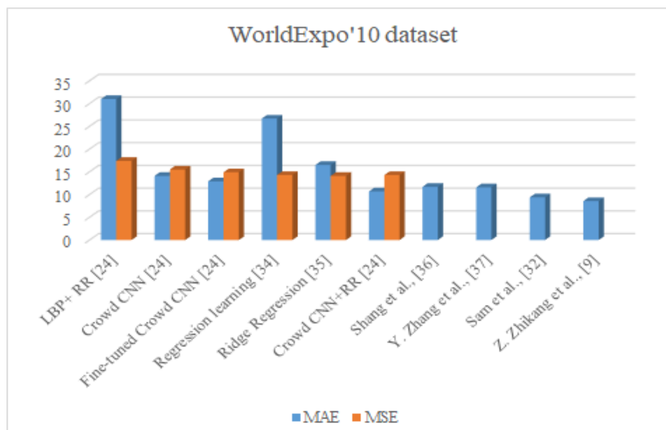
TABLE. VI. COMPARATIVE STUDIES BASED ON FOUND RESULTS RELATED TO MAKE3D DATASET

Methods	Root Mean Squared
Walach et al., [17]	13.89
Saxena et al., [54]	16.7
Li et al., [55]	15.2
Karch et al., [56]	15.1
F. Liu et al., [57]	12.89
M. Liu et al., [58]	12.6

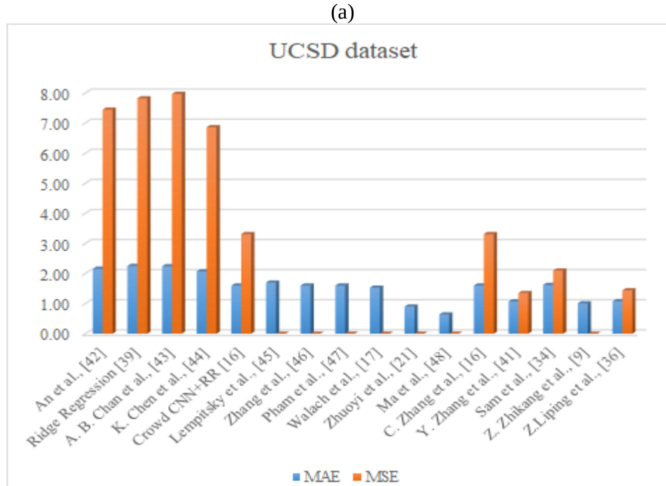
TABLE. VII. COMPARATIVE STUDIES BASED ON FOUND RESULTS RELATED TO SHANGHAITECH (PART A) DATASET

Methods	MAE	MSE
Z. Zhikang et al., [9]	17.5	103.5
LBP+ RR [16]	59.1	303.2
C. Zhang et al., [16]	49.8	336.1
Sam et al., [34]	21.6	135.0
Z.Liping et al., [36]	65.6	107.2
X. Zeng et al., [37]	68.2	107.8
Y. Zhang et al., [41]	26.4	173.2
Sindagi et al., [52]	20.0	152.4

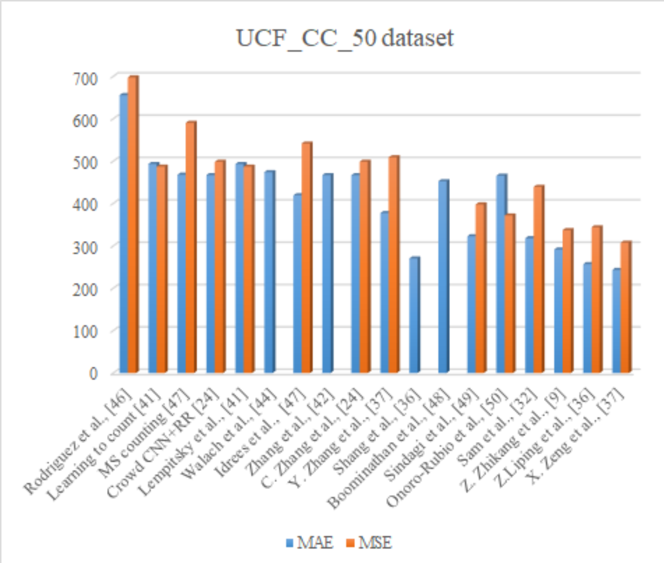
Fig. 1 highlights the difference between related works according to the MAE and MSE.



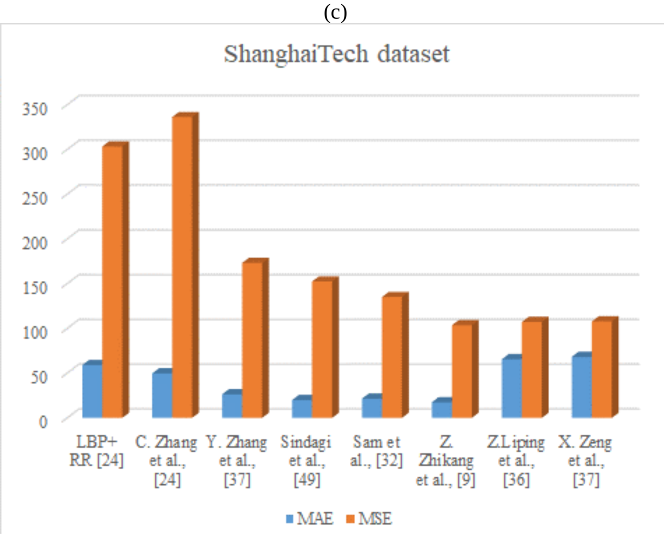
(a)



(b)



(c)



(d)

Fig. 1. A Comparison between Related Works According to different Datasets.

VI. Conclusions

The estimation of the density of the crowded area is still a challenge for researchers. In this paper, we have presented the different research axes related to the crowded zone especially in terms of the datasets, metrics, and approaches.

Datasets should be improved by increasing the number of scenes and achieve over than half-million pedestrians. These enhancements are required to improve results found in the evaluation phase.

Approaches still request improvement by decreasing the error between the real number of crowd and the estimated counts. Other methods would be applied to support the movement of the camera and the fusion of data received from many sources. In addition, the Real-Time constraint has to be the future of research work related to this domain.

The estimation of the density of crowd area is employed in a different type of applications as the estimation of pedestrians, crowded car traffic, the crowd in malls, and bacterial cell microscopy.

Besides, this domain continues to be an open challenge according to the number of the article published in the Web of Science database. Publishing is boosting from 85 articles in 2015 to achieve 148 articles published in 2019. The IEEE Xplore database shows that the number of published papers between 2018 and 2019 is increased by 25%. In the Science direct database, about 20% of the evolution of published papers are highlighted since 2018.

These statistics prove not only the importance of the domain but also the continuity of the challenge facing the crowd's estimation.

According to this deep study, the crowd size's estimation still requests enhancement in accuracy and real-time constraints.

References

[1] M. Ben Ayed, S. Elkosantini, and M. Abid, "An Automated Surveillance System Based on Multi-Processor and GPU Architecture," (in English), Engineering Technology & Applied Science Research, Article vol. 7, no. 6, pp. 2319-2323, Dec 2017.

[2] K. A. Eldrandaly, M. Abdel-Basset, and L. Abdel-Fatah, "PTZ-Surveillance coverage based on artificial intelligence for smart cities," (in English), International Journal of Information Management, Article vol. 49, pp. 520-532, Dec 2019.

[3] R. Arroyo, J. J. Yebes, L. M. Bergasa, I. G. Daza, and J. Almazan, "Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls," (in English), Expert Systems with Applications, Article vol. 42, no. 21, pp. 7991-8005, Nov 2015.

[4] J. J. Feng and Y. He, "Collective motion of bacteria and their dynamic assembly behavior," (in English), Science China-Materials, Article; Proceedings Paper vol. 60, no. 11, pp. 1079-1092, Nov 2017.

[5] J. C. Chen et al., "A Cascaded Convolutional Neural Network or Age Estimation of Unconstrained Faces," in 2016 Ieee 8th International Conference on Biometrics Theory, Applications and Systems(International Conference on Biometrics Theory Applications and Systems, New York: Ieee, 2016.

[6] A. Sieben, J. Schumann, and A. Seyfried, "Collective phenomena in crowds-Where pedestrian dynamics need social psychology," (in English), Plos One, Article vol. 12, no. 6, p. 19, Jun 2017, Art. no. e0177328.

[7] M. B. Ayed, S. Elkosantini, S. A. Alshaya, and M. Abid, "Suspicious Behavior Recognition Based on Face Features," IEEE Access, vol. 7, pp. 149952-149958, 2019.

[8] M. Ben Ayed, SabeurElkosantini, and M. Abid, "An Automated Surveillance System based on Multi-Processor System-on-Chip and Hardware Accelerator," (in English), International Journal of Advanced Computer Science and Applications, Article vol. 8, no. 9, pp. 59-66, Sep 2017.

[9] Z. K. Zou, Y. Cheng, X. Y. Qu, S. L. Ji, X. X. Guo, and P. Zhou, "Attend to count: Crowd counting with adaptive capacity multi-scale CNNs," (in English), Neurocomputing, Article vol. 367, pp. 75-83, Nov 2019.

[10] A. Mahmood and S. Al-Maadeed, "Action recognition in poor-quality spectator crowd videos using head distribution-based person segmentation," (in English), Machine Vision and Applications, Article vol. 30, no. 6, pp. 1083-1096, Sep 2019.

[11] H. H. Chen, B. Guo, Z. W. Yu, and Q. Han, "CrowdTracking: Real-Time Vehicle Tracking Through Mobile Crowdsensing," (in English), Ieee Internet of Things Journal, Article vol. 6, no. 5, pp. 7570-7583, Oct 2019.

[12] Y. Ma, E. W. Lee, Z. A. Hu, M. Shi, and R. K. Yuen, "An Intelligence-Based Approach for Prediction of Microscopic Pedestrian Walking

Behavior," (in English), Ieee Transactions on Intelligent Transportation Systems, Article vol. 20, no. 10, pp. 3964-3980, Oct 2019.

[13] X. L. Wei, J. P. Du, M. Y. Liang, and L. F. Ye, "Boosting deep attribute learning via support vector regression for fast moving crowd counting," (in English), Pattern Recognition Letters, Article vol. 119, pp. 12-23, Mar 2019.

[14] A. S. Rao, J. Gubbi, S. Marusic, and M. Palaniswami, "Estimation of crowd density by clustering motion cues," (in English), Visual Computer, Article vol. 31, no. 11, pp. 1533-1552, Nov 2015.

[15] K. Aziz, D. Merad, R. Iguernaissi, P. Drap, and B. Fertil, "Head detection based on skeleton graph method for counting people in crowded environments," (in English), Journal of Electronic Imaging, Article vol. 25, no. 1, p. 14, Jan 2016, Art. no. 013012.

[16] C. Zhang, H. S. Li, X. G. Wang, X. K. Yang, and Ieee, "Cross-scene Crowd Counting via Deep Convolutional Neural Networks," in 2015 Ieee Conference on Computer Vision and Pattern Recognition(IEEE Conference on Computer Vision and Pattern Recognition, New York: Ieee, 2015, pp. 833-841.

[17] E. Walach and L. Wolf, "Learning to count with cnn boosting," in European Conference on Computer Vision, 2016, pp. 660-676: Springer.

[18] E. Walach and L. Wolf, "Learning to Count with CNN Boosting," in Computer Vision - Eccv 2016, Pt Ii, vol. 9906, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. (Lecture Notes in Computer Science, Cham: Springer International Publishing Ag, 2016, pp. 660-676.

[19] Y. Y. Zhang, D. S. Zhou, S. Q. Chen, S. H. Gao, Y. Ma, and Ieee, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," in 2016 Ieee Conference on Computer Vision and Pattern Recognition(IEEE Conference on Computer Vision and Pattern Recognition, New York: Ieee, 2016, pp. 589-597.

[20] S. Jachner, G. Van den Boogaart, and T. Petzoldt, "Statistical methods for the qualitative assessment of dynamic models with time delay (R Package qualV)," Journal of Statistical Software, vol. 22, no. 8, pp. 1-30, 2007.

[21] Z. Y. Zhao, H. S. Li, R. Zhao, and X. G. Wang, "Crossing-Line Crowd Counting with Two-Phase Deep Neural Networks," in Computer Vision - Eccv 2016, Pt Viii, vol. 9912, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. (Lecture Notes in Computer Science, Cham: Springer International Publishing Ag, 2016, pp. 712-726.

[22] S. F. Lin, J. Y. Chen, and H. X. Chao, "Estimation of number of people in crowded scenes using perspective transformation," (in English), Ieee Transactions on Systems Man and Cybernetics Part a-Systems and Humans, Article vol. 31, no. 6, pp. 645-654, Nov 2001.

[23] A. C. Davies, J. H. Yin, and S. A. Velastin, "CROWD MONITORING USING IMAGE-PROCESSING," (in English), Electronics & Communication Engineering Journal, Article vol. 7, no. 1, pp. 37-47, Feb 1995.

[24] C. S. Regazzoni and A. Tesei, "Distributed data fusion for real-time crowding estimation," (in English), Signal Processing, Article vol. 53, no. 1, pp. 47-63, Aug 1996.

[25] A. N. Marana, L. D. Costa, R. A. Lotufo, S. A. Velastin, Ieee, and Ieee, "Estimating crowd density with Minkowski fractal dimension," in Icassp '99: 1999 Ieee International Conference on Acoustics, Speech, and Signal Processing, Proceedings Vols I-Vi(International Conference on Acoustics Speech and Signal Processing (ICASSP), New York: Ieee, 1999, pp. 3521-3524.

[26] R. M. Haralick, "STATISTICAL AND STRUCTURAL APPROACHES TO TEXTURE," (in English), Proceedings of the Ieee, Review vol. 67, no. 5, pp. 786-804, 1979.

[27] S. Y. Cho, T. W. S. Chow, and C. T. Leung, "A neural-based crowd estimation by hybrid global learning algorithm," (in English), Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics, Article vol. 29, no. 4, pp. 535-541, Aug 1999.

[28] C. Wang, H. Zhang, L. Yang, S. Liu, X. C. Cao, and Acm, Deep People Counting in Extremely Dense Crowds (Mm'15: Proceedings of the 2015 Acm Multimedia Conference). New York: Assoc Computing Machinery, 2015, pp. 1299-1302.

[29] M. Fu, P. Xu, X. D. Li, Q. H. Liu, M. Ye, and C. Zhu, "Fast crowd density estimation with convolutional neural networks," (in English), Engineering Applications of Artificial Intelligence, Article vol. 43, pp. 81-88, Aug 2015.

[30] C. Shang, H. Z. Ai, B. Bai, and Ieee, "END-TO-END CROWD COUNTING VIA JOINT LEARNING LOCAL AND GLOBAL COUNT," in 2016 Ieee International Conference on Image Processing(IEEE International Conference on Image Processing ICIP, New York: Ieee, 2016, pp. 1215-1219.

[31] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, "A Large Contextual Dataset for Classification, Detection and Counting of Cars with Deep Learning," in Computer Vision - Eccv 2016, Pt Iii, vol. 9907, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. (Lecture Notes in Computer Science, Cham: Springer Int Publishing Ag, 2016, pp. 785-800.

[32] L. Boominathan, S. S. S. Kruthiventi, R. V. Babu, and Acm, CrowdNet: A Deep Convolutional Network for Dense Crowd Counting (Mm'16: Proceedings of the 2016 Acm Multimedia Conference). New York: Assoc Computing Machinery, 2016, pp. 640-644.

[33] V. A. Sindagi, V. M. Patel, and Ieee, CNN-based Cascaded Multi-task Learning of High-level Prior and Density Estimation for Crowd Counting (2017 14th Ieee International Conference on Advanced Video and Signal Based Surveillance). New York: Ieee, 2017.

[34] D. B. Sam, S. Surya, R. V. Babu, and Ieee, "Switching Convolutional Neural Network for Crowd Counting," in 30th Ieee Conference on Computer Vision and Pattern Recognition(IEEE Conference on Computer Vision and Pattern Recognition, New York: Ieee, 2017, pp. 4031-4039.

[35] X. X. Yang, X. L. Yang, Q. L. Wang, Y. L. Kang, and F. Q. Pan, "Guide optimization in pedestrian emergency evacuation," (in English), Applied Mathematics and Computation, Article vol. 365, p. 12, Jan 2020, Art. no. Unsp 124711.

[36] L. Zhu, C. Li, B. Wang, K. Yuan, and Z. Yang, "DCGSA: A global self-attention network with dilated convolution for crowd density map generating," Neurocomputing, 2019.

[37] X. Zeng, Y. Wu, S. Hu, R. Wang, and Y. Ye, "DSPNet: Deep scale purifier network for dense crowd counting," Expert Systems with Applications, vol. 141, p. 112977, 2020.

[38] L. Fiaschi, U. Köthe, R. Nair, and F. A. Hamprecht, "Learning to count with regression forest and structured labels," in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), 2012, pp. 2685-2688: IEEE.

[39] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in BMVC, 2012, vol. 1, no. 2, p. 3.

[40] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 1215-1219: IEEE.

[41] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 589-597.

[42] S. An, W. Liu, and S. Venkatesh, "Face recognition using kernel ridge regression," in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-7: IEEE.

[43] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-7: IEEE.

[44] K. Chen, S. Gong, T. Xiang, and C. Change Loy, "Cumulative attribute space for age and crowd density estimation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2467-2474.

[45] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in Advances in neural information processing systems, 2010, pp. 1324-1332.

[46] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 833-841.

[47] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3253-3261.

[48] Z. Ma and A. B. Chan, "Crossing the line: Crowd counting by integer programming with local features," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2539-2546.

[49] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert, "Density-aware person detection and tracking in crowds," in 2011 International Conference on Computer Vision, 2011, pp. 2423-2430: IEEE.

[50] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2547-2554.

[51] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 640-644: ACM.

[52] V. A. Sindagi and V. M. Patel, "Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1-6: IEEE.

[53] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in European Conference on Computer Vision, 2016, pp. 615-629: Springer.

[54] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in Advances in neural information processing systems, 2006, pp. 1161-1168.

[55] C. Li, A. Kowdle, A. Saxena, and T. Chen, "Towards holistic scene understanding: Feedback enabled cascaded classification models," in Advances in Neural Information Processing Systems, 2010, pp. 1351-1359.

[56] K. Karsch, C. Liu, and S. B. Kang, "Depth transfer: Depth extraction from video using non-parametric sampling," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 11, pp. 2144-2158, 2014.

[57] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 10, pp. 2024-2039, 2015.

[58] M. Liu, M. Salzmann, and X. He, "Discrete-continuous depth estimation from a single image," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 716-723.

Model for Measuring benefit of Government IT Investment using Fuzzy AHP

Prih Haryanta[1], Azhari Azhari[2], Khabib Mustofa[3]
Department of Computer Science and Electronics
Faculty of Mathematic and Natural Sciences
Universitas Gadjah Mada, Yogyakarta, Indonesia

***Abstract*—Information Technology (IT) has become a mandatory for every organization including government. Investment on IT can help government to deliver services to the citizen. Every IT investment should give the maximum result. Measurement for the benefit of IT investment is needed to make sure that it has deliver the missions and goals. There are plenty models for measuring the feasibility of an IT investment before the implementation. But there are still few models to measure the IT investment after implementation. This paper proposes a model to measure the benefit of an IT investment after implementation, especially in government organizations. The model uses generic IS/IT business value category which consists of 13 categories and 73 sub-categories. Each category will be weighted according to organization preference using Fuzzy Analitic Hierarchy Process (FAHP). This model is applied to measure IT investments in the Ministry of Finance of the Republic of Indonesia, named SPAN and SAKTI applications. The weighted benefit score of SPAN is 76.39%, while the original score is 75.89%. The weighted benefit score of SAKTI is 68.08%, while the original score is 67.33%. The differences between the original score and weighted score indicate that the model accommodates the organization's preference in the evaluation.**

Keywords—IT investment; government investment; ex-post evaluation; benefit creation; fuzzy AHP; analytic hierarchy process

I. Introduction

The use of Information Technology (IT) will increase the atmosphere of openness and transparency [1]. Thus, IT can be used by the Government to deliver transparent and accountable services. Investment in IT by Government is needed to support the organization. The IT investment that has been implemented by the government needs to be evaluated as an embodiment of accountable governance. According to the IT governance framework, The Control Objectives for Information and Related Technology (COBIT) version 5, in the EDM02 process -Ensure Benefits Delivery, organizations must be able to ensure that IT benefits have actually been achieved and are received by all stakeholders.

After an investment being implemented, there is a need to measure that the benefit of investment has been delivered to the stakeholder. The measurement problem is a significant factor that becomes an obstacle in evaluation [2][3]. It is difficult to identify the benefits in ex-post evaluations [4]. That is because of the different systems used by various business areas so that the benefits of each business area can be varied. The evaluation is complicated because the definition of success is unclear and varies from one organization to another [5]. The unclear definition of benefit/value makes an imbalance between theory and practice [6].

Ex-post evaluation models for government IT investments are still very limited in number. Some ex-post IT investment evaluation models used by governments in the world like Social Return on Investment (SROI), Balanced E-Government Index (BEGIX), Public Sector Value Model (PSV), Performance Reference Model (PRM), Interchange of Data between Administration Value of Investment (IDA VOI), Method of Analysis and Value Enhancement (MAREVA), E-Gov Economics Projects (eGEP) [7][8]. Some of the above models (SROI, BEGIX, IDA VOI, MAREVA) use financial-based to measure the benefit/value creation. While for government institutions, it is not always about financial benefit. A literature review on Information and Communication Technologies (ICT) project evaluation by AL-Ghamdi et al [9] found that the post implementation evaluation approach of the ICT project in common practice are assessing non-financial ICT business values. The PRM model focuses more on organizational performance in general and not on the impact of IT investment results. Although the PSV model does not measure the value of money, it does not show the specific value obtained from the IT investment. While the eGEP model is very broad in scope, its measurement is not suitable for measuring a single IT investment by the government. Setiawan et al [10] proposed a hybrid method for evaluating the performance of ICT projects. Although the methods can be used to evaluate the performance of ICT projects, it cannot determine what benefits are created from the ICT projects being evaluated.

Therefore, there is a need to develop a new ex-post evaluation method that can show the benefit/value of an IT investment, especially in government organizations. It should overcome the problem of benefit differences in various organizations. Also, the method should accommodate the organization's preference in the evaluation. This study proposed a model for identifying benefit creation of an IT investment in government organizations. The benefit created will be scored and weighted according to decision makers in the organization.

This paper is divided into six sections: the first section is the introduction, the second and third sections are the literature review, the fourth section is the research methodology, the fifth sections present the result and discussion, and the last section present the conclusions generated from this research.

This research was supported by the Ph.D. grant of the Indonesia endowment fund for education (LPDP), Ministry of Finance of The Republic of Indonesia.

II. Generic IS/IT Business Value Template

The generic IS/IT Business value template was proposed by Ranti [11]. The IS/IT business value template consists of 13 benefit categories and 73 benefit sub-categories, they are:

1) Reducing Cost of (traveling cost, staff/operator/ employee cost, meeting cost, service failure cost, application development cost, delivery cost, training cost per employee, returning cost for incorrect delivery, cost of money, office supplies and printing cost, subscription cost of certain reading materials or subscription cost per employee, space rental cost, device rental cost, inventory cost, research failure cost).

2) Increasing Productivity caused by (restructuring job function, accelerating mastering product knowledge, ease of analysis, increasing employee satisfaction).

3) Accelerating Process of (production process, stock procurement process, report making process, data preparation process, order checking process, debt payment process, transaction process, decision-making process).

4) Reducing Risk of (price miscalculation, unrecoverable claim, inventory lost, rejected goods, data lost, incorrect data, penalty, losing potential employee, forgery, administration fraud, incorrect payment, asset mismanagement).

5) Increasing Revenue caused by (increasing business capacity, increasing report quality, increasing customer trust, widening market segment, increasing other incomes).

6) Increasing Accuracy of (billing, analysis, data, planning, decision).

7) Accelerating Cash-in caused by (accelerating billing dispatching).

8) Increasing External Services of (reducing order cancellation, knowing customer's problems, adding a point of services, personalized services, customer satisfaction).

9) Increasing Image caused by (increasing service quality, offering substantial discounts, complying with regulations, using branded systems).

10) Increasing Quality of (better supplier/vendor management, work result, services, products).

11) Increasing Internal Services of (shared services, matching employee's right and responsibility, employee services, proper schedule, and training material).

12) Increasing Competitive Advantage caused by (forming business alliances, accelerating the execution of new business opportunities, increasing switching costs).

13) Avoiding Cost of (reserved fund, maintenance cost, lost and delay cost).

This template could be used to overcome the problem of benefit differences in various organizations. The generic IS/IT Business value template will be used as a base for measurement.

III. Fuzzy AHP

Analytic Hierarchy Process (AHP) is a method proposed by Saaty [12][16] for selecting alternatives using distinct criteria. The AHP method breaks a complex and unstructured problem into several components in a hierarchical arrangement. The decision-maker makes a pairwise comparison between criteria. This method can be used for weighting the criteria for decision making. AHP deals with a crisp number to represents the judgment by the decision-maker. The intensity scale of importance converted into a number for computation.

Fuzzy AHP (FAHP) method proposed by Laarhoven and Pedrycz [13] is an extension of AHP using fuzzy concept. The scale of importance is represented in fuzzy using Triangular Fuzzy Number (TFN) (Table I). FAHP could deal with subjective judgment in making priorities. Sehra et al. demonstrated the different results of software quality model selection using AHP and FAHP [14].

Using Extend Analysis by Chang [15], Fuzzy AHP technique is divided into several steps:

1) Creating a pairwise comparison matrix between categories in the TFN scale.

2) Calculating the fuzzy synthetic extents ($\tilde{S}_x$) of the above matrix on category *x* with the equation:

$$(\tilde{S}_x = \sum_{y=1}^{n} \tilde{C}_{xy} \otimes \left[\sum_{k=1}^{n}\sum_{y=1}^{n} \tilde{C}_{ky}\right]^{-1}; \text{x=1, 2, ..., n} \quad (1)$$

Where $\otimes$ denotes the extended multiplication of two fuzzy numbers, n is the size of the pairwise comparison matrix between categories, and k is a combination of criteria from line i where i = 1 to n.

$$\sum_{y=1}^{n} \tilde{C}_{xy} = \left(\sum_{y=1}^{n} l_{xy}, \sum_{y=1}^{n} m_{xy}, \sum_{y=1}^{n} u_{xy}\right); \text{x=1, 2, ..., n} \quad (2)$$

where l is the lower bound, m is the middle bound, u is the upper bound.

$$\left[\sum_{k=1}^{n}\sum_{y=1}^{n} \tilde{C}_{ky}\right]^{-1} = \left[\frac{1}{\sum_{k=1}^{n}\sum_{y=1}^{n} u_{ky}}, \frac{1}{\sum_{k=1}^{n}\sum_{y=1}^{n} m_{ky}}, \frac{1}{\sum_{k=1}^{n}\sum_{y=1}^{n} l_{ky}}\right] \quad (3)$$

$$\sum_{k=1}^{n}\sum_{y=1}^{n} \tilde{C}_{ky} = \left(\sum_{k=1}^{n}\sum_{y=1}^{n} l_{ky}, \sum_{k=1}^{n}\sum_{y=1}^{n} m_{ky}, \sum_{k=1}^{n}\sum_{y=1}^{n} u_{ky}\right) = \left[\left(\sum_{y=1}^{n} l_{1y}, \sum_{y=1}^{n} m_{1y}, \sum_{y=1}^{n} u_{1y}\right) + \cdots + \left(\sum_{y=1}^{n} l_{ny}, \sum_{y=1}^{n} m_{ny}, \sum_{y=1}^{n} u_{ny}\right)\right] \quad (4)$$

3) Comparing the fuzzy synthetic extents ($\tilde{S}_x$) of one category with another fuzzy synthetic extents category ($\tilde{S}_y$), which is called as degree of possibility with equation:

$$V(\tilde{S}_x \geq \tilde{S}_y) \begin{cases} 0, \text{if } m_x \geq m_y \\ 1, \text{if } l_y \geq u_x \\ \frac{l_y - u_x}{(m_x - u_x) - (m_y - l_y)}, \text{otherwise} \end{cases} \quad (5)$$

Where $V(\tilde{S}_x \geq \tilde{S}_y \mid \text{y} = 1, \ldots, \text{n}); \text{y} \neq \text{x}$

4) Determining the minimum degree of possibility of $V(\tilde{S}_x \geq \tilde{S}_y)$

$$d'(A_x) = \min V(\tilde{S}_x \geq \tilde{S}_y) \quad (6)$$

Then the weight vector (W') is given by

$$W' = \left(d'(A_1), d'(A_2), \ldots, d'(A_n)\right)^T \quad (7)$$

Where $A_1 (i = 1,2, \ldots, n)$ are n elements.

TABLE. I. TRIANGULAR FUZZY NUMBER

Linguistic Variable	Positive TFN	Positive Reciprocal TFN
Extremely Strong (ES)	(9,9,9)	(1/9,1/9,1/9)
Intermediate value	(6,8,9)	(1/9,1/8,1/6)
Very Strong (VS)	(5,7,9)	(1/9,1/7,1/5)
Intermediate value	(4,6,8)	(1/8,1/6,1/4)
Strong (S)	(3,5,7)	(1/7,1/5,1/3)
Intermediate value	(2,4,6)	(1/6,1/4,1/2)
Moderately Strong (MS)	(1,3,5)	(1/5,1/3,1)
Intermediate value	(1,2,4)	(1/4,1/2,1)
Equally Strong (EqS)	(1,1,1)	(1,1,1)

5) Then the normalized importance weight vector W of the *pairwise comparison* are:

$$W = \left(d(A_1), d(A_2), \dots, d(A_n)\right)^T \tag{8}$$

where W is a nonfuzzy number that represents the priority weights of an attribute or an alternative over another.

IV. RESEARCH METHODOLOGY

The proposed model is for measuring the benefit generated by an IT investment in government organizations. To identify the benefit, this model is using the Generic IS/IT Business Value Template. The use of the template is intended to overcome the problem of benefit differences in various organizations. Secondly, this model will use Fuzzy AHP to weighting the benefit category. The weighting process is needed to accommodate the organization's preference.

The first step of the proposed model is to determine the categories and subcategories of business benefits that are created with IT investment. This step will be carried out by distributing questionnaires to stakeholders involved in implementing IT investments. With this questionnaire method, we will get benefit categories and subcategories that have contributed to the creation of benefit with IT investment. In the survey respondents are choosing the benefit creation from each benefit sub-category based on their experiences. Respondents fill based on personal experience in the field by following intuition, experience, data, information and critical level possessed by the assessor [17]. Filling in the benefit creation by using a scale in lingual form with the following percentage ranges:

a) Very high benefits ((80-100%])

b) High benefits ((60-80%])

c) Medium benefits ((40-60%])

d) Low benefits ((20-40%])

e) Very low benefits ((0-20%])

f) No benefits created (0%)

The results of the questionnaire then processed to obtain the benefit score. The score of benefit categories is the average score of the benefit sub-categories from particular categories. The score of each benefit categories then multiplied by the weight of the categories. The weight of each benefit categories is made by decision-maker using pairwise comparison between the categories. The result of comparison is then processed by the Fuzzy AHP method to get the importance weight of each benefit category. The final score of benefit creation is the sum of the weighted values of each benefit category.

To validate the proposed model, the model has been used to evaluating two IT investment in the Ministry of Finance of the Republic of Indonesia (MoF), i.e.

1) Sistem Perbendaharaan dan Anggaran Negara (SPAN) (State Treasury and Budget Application System).

2) Sistem Aplikasi Keuangan Tingkat Instansi (SAKTI) (Institution-level Financial Application System).

SPAN is an IT investment made by MoF to integrate and centralize the financial management information system in Indonesia. It replaced the old system which was distributed across the country. SAKTI is also an IT investment made by MoF to integrate many financial application in the operating ministries. SPAN and SAKTI application now support the new financial system for the Government of the Republic of Indonesia.

V. RESULT AND DISCUSSION

To collect the opinion from the stakeholder, questionnaire has been given to the user to measure the benefit creation based on their experience. The questionnaire is based on the Generic IS/IT Business value template. For weighting the category, the decision-maker in the organization makes a pairwise comparison between the benefit categories. There are 13 benefit categories, i.e reducing cost (C1), increasing productivity (C2), accelerating process (C3), reducing risk (C4), increasing revenue (C5), increasing accuracy (C6), accelerating cash-in (C7), increasing external services (C8), increasing image (C9), increasing quality (C10), increasing internal services (C11), increasing competitive advantage (C12), and avoiding cost (C13). The result of the pairwise comparison in the TFN scale is presented in Table II.

Using extent analysis by Chang (equation 1 to 8), the result of importance weight vectors after normalization is presented in Table III. The weight value of the benefit category summarized and presented in Table IV. The weight of the benefit category then used for calculating the total benefit created from the IT investment.

The questionnaire was given to the SPAN and SAKTI applications users. There are 30 respondents for each application. The score of each benefit category is obtained from the average score of sub-categories benefit in the same category. The benefit score then multiplied by the weight of the benefit category. The final benefit score is the sum of the weighted score of all benefit categories. The result is shown in Table V and Table VI.

The result of SPAN investment, the total weighted benefit score (76.39%) is higher than the original score (75.89%). This indicates that some benefit categories with high priority has a higher score among others. On the other hand, the top priority benefit category (reducing risk) has a score of 78.11% that categorized as a high benefit. And the least score of benefit category belongs to benefit "increasing competitive advantage" which is the 7th priority out of 13.

TABLE. II. PAIRWISE COMPARISON IN TFN

	C1			C2			C3			C4			C5			C6			C7			C8			C9			C10			C11			C12			C13		
	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u
C1	1	1	1	1	3	5	1	3	5	1	3	5	1	1	1	1	1	1	1	3	5	1	1	1	1	3	5	1	1	1	1	3	5	1	1	1	1	1	1
C2	0,2	0,33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	5	1	1	1	1	1	1	1	1	1	1	3	5	1	1	1
C3	0,2	0,33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5
C4	0,2	0,33	1	1	1	1	1	1	1	1	1	1	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5
C5	1	1	1	1	1	1	1	1	1	0,2	0,33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	5	1	1	1	1	3	5
C6	1	1	1	1	1	1	1	1	1	0,2	0,33	1	1	1	1	1	1	1	1	3	5	1	1	1	1	1	1	1	1	1	1	1	1	1	3	5	1	3	5
C7	0,2	0,33	1	1	1	1	0,2	0,33	1	0,2	0,33	1	1	1	1	0,2	0,33	1	1	1	1	1	1	1	1	3	5	1	1	1	1	1	1	5	7	9	1	1	1
C8	1	1	1	0,2	0,33	1	0,2	0,33	1	0,2	0,33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C9	0,2	0,33	1	1	1	1	0,2	0,33	1	0,2	0,33	1	1	1	1	1	1	1	0,2	0,33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C10	1	1	1	1	1	1	0,2	0,33	1	0,2	0,33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C11	0,2	0,33	1	1	1	1	0,2	0,33	1	0,2	0,33	1	0,2	0,33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C12	1	1	1	0,2	0,33	1	0,2	0,33	1	0,2	0,33	1	1	1	1	0,2	0,33	1	0,11	0,14	0,2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	5
C13	1	1	1	1	1	1	0,2	0,33	1	0,2	0,33	1	0,2	0,33	1	0,2	0,33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,2	0,33	1	1	1	1

TABLE. III. NORMALIZED IMPORTANCE WEIGHT

	d(A1)	d(A2)	d(A3)	d(A4)	d(A5)	d(A6)	d(A7)	d(A8)	d(A9)	d(A10)	d(A11)	d(A12)	d(A13)
W	0,1234	0,0842	0,1279	0,1379	0,0478	0,0948	0,0870	0,0478	0,0468	0,0489	0,0468	0,0610	0,0457

TABLE. IV. WEIGHT OF BENEFIT CATEGORY

	Benefit Category	**Weight**
C1	Reducing Cost	0,1234
C2	Increasing Productivity	0,0842
C3	Accelerating Process	0,1279
C4	Reducing Risk	0,1379
C5	Increasing Revenue	0,0478
C6	Increasing Accuracy	0,0948
C7	Accelerating Cash-in	0,0870
C8	Increasing External Services	0,0478
C9	Increasing Image	0,0468
C10	Increasing Quality	0,0489
C11	Increasing Internal Services	0,0468
C12	Increasing Competitive Advantage	0,0610
C13	Avoiding Cost	0,0457

TABLE. V. SPAN BENEFIT SCORE

	Benefit Category	**Score**	**Weight**	**Weighted Score**
1	Reducing Cost	72,28	0,1234	8,9192
2	Increasing Productivity	77,00	0,0842	6,4848
3	Accelerating Process	78,67	0,1279	10,0629
4	Reducing Risk	78,11	0,1379	10,7722
5	Increasing Revenue	71,48	0,0478	3,4175
6	Increasing Accuracy	80,90	0,0948	7,6684
7	Accelerating Cash-in	79,83	0,0870	6,9468
8	Increasing External Services	75,16	0,0478	3,5933
9	Increasing Image	76,32	0,0468	3,5682
10	Increasing Quality	79,00	0,0489	3,8639
11	Increasing Internal Services	76,00	0,0468	3,5534
12	Increasing Competitive Advantage	68,92	0,0610	4,2016
13	Avoiding Cost	72,94	0,0457	3,3370
	Total	**75,89**[a]		**76,39**

[a]average

TABLE. VI. SAKTI BENEFIT SCORE

	Benefit Category	Score	Weight	Weighted Score
1	Reducing Cost	63,43	0,1234	7,8275
2	Increasing Productivity	68,98	0,0842	5,8097
3	Accelerating Process	71,64	0,1279	9,1643
4	Reducing Risk	71,03	0,1379	9,7958
5	Increasing Revenue	57,75	0,0478	2,7611
6	Increasing Accuracy	75,55	0,0948	7,1610
7	Accelerating Cash-in	66,43	0,0870	5,7807
8	Increasing External Services	62,91	0,0478	3,0077
9	Increasing Image	64,36	0,0468	3,0093
10	Increasing Quality	70,05	0,0489	3,4260
11	Increasing Internal Services	70,98	0,0468	3,3185
12	Increasing Competitive Advantage	64,00	0,0610	3,9015
13	Avoiding Cost	68,13	0,0457	3,1167
	Total	**67,33**[b]		**68,08**

[b]average

The similar results gained for SAKTI investment. The total weighted benefit score (68.08%) is higher than the original score (67.33%). It also indicates that some benefit categories with high priority has a higher score among others. The top priority benefit category (reducing risk) got score 71.03% that categorized as high benefit. While the least score of benefit category belongs to benefit “increasing revenue” with score of 57.75% that categorized as a medium benefit.

From those two IT investments, the final score of the benefit creation is categorized as “high benefits”. This category still can be improved to become a “very high benefits” category. The leader of the organization could use the results of these measurements as a base for evaluation to improve the achievement of IT investment benefits.

VI. CONCLUSION AND FUTURE WORK

Based on the result and discussion, the proposed model could fulfill the research objectives. First, the model can overcome the problem of differences in the benefits criteria between organizations and can measure the benefits created from an IT investment in government organizations. Second, the differences between the score using FAHP weighting and the original score indicates that the model accommodates the organization’s preference in the evaluation. Third, the proposed model can be a supporting tool to meet the requirements of COBIT 5 framework, especially in the EDM02 process - Ensure Benefits Delivery.

For future work, researchers can develop new business value templates that are better suited to the nature of government organizations. In addition, researchers can also try other weighting methods and compare the results with the results of the Fuzzy AHP method.

ACKNOWLEDGMENT

The authors would like to thank all the respondents and the decision maker in the Ministry of Finance for their time and effort to help complete this research.

REFERENCES

[1] Bertot, J. C., Jaeger, P. T., and Grimes, J. M., “Using ICTs to create a culture of transparency : E-government and social media as openness and anti-corruption tools for societies.” Government Information Quarterly 27 (2010) pp. 264-271

[2] Willcocks, L., and Lester, S., "Beyond the IT Productivity Paradox," European Management Journal Vol. 14, No.3, pp. 279-290, 1996.

[3] Ballantine, J.A., Galliers, R.D., and Stray, S.J., "Information system/technology evaluation practices: evidence from UK organizations," Journal of Information Technology (1996) 11, 129-141.

[4] Norris, G., "Post-investment Appraisal - Investing in Information Systems: Evaluation and Management," Chapman and Hall, London, 1996.

[5] Hillam, C., Edwards, H., and Young, C., "Company approaches to IT/IS investment and the resulting evaluation," Seventh European Conference on Information Technology Evaluation, Brown, A., and Remenyi, D. (eds), Trinity College, Dublin, Ireland, pp.109-116, 2000.

[6] Bannister, F., and Remenyi, D., "Value perception in IT Investment Decisions," The Electronic Journal of Information Systems Evaluation, Vol. 2, No. 2, 1999.

[7] Dadayan, L., "Measuring Return on Government IT Investments," Proceeding of European Conference on Information Technology Evaluation, Genoa, 2006.

[8] Palka, W., Gülümser, Y., Jurisch, M.C., Wolf, P. dan Krcmar, H., "Basic Approaches for the Evaluation of IT-Investments in E-Government: A Literature Review," Electronic Government and Electronic Participation - Joint Proceedings of Ongoing Research of IFIP EGOV and IFIP ePart 2014, Gesellschaft für Informatik e.V., Bonn, pp. 27-34, 2014.

[9] AL-Ghamdi, A.S.A., Saleem, F., "General Characteristics and Common Practices for ICT Projects: Evaluation Perspective", International Journal of Advanced Computer Science and Applications Vol.9 No.1, 2018.

[10] Setiawan, H., Istiyanto, J.E., Wardoyo, R., Santoso, P., "The Group Decision Support System To Evaluate The ICT Project Performance Using The Hybrid Method Of AHP, TOPSIS And Copeland Score", International Journal of Advanced Computer Science and Applications Vol.7 No.4, 2016.

[11] Ranti, B., "The Generic IS/IT Business Value Category : Cases in Indonesia," Konferensi dan Temu Nasional TIK Untuk Indonesia, Jakarta, 2008.

[12] Saaty, T.L., "A Scaling method for priorities in hierarchichal structure," Journal of Mathematical Psychology, 15, pp. 234-281, 1977.

[13] Van Laarhoven, P.J.M. dan Pedrycz, W., "A Fuzzy Extension of Saaty's Priority Theory," Fuzzy Sets and Systems (11), pp.229-241, 1983.

[14] Sehra, S.K., Brar, Y.S., Kaur, N., "Application of Multi-Criteria Decision Making in Software Engineering", International Journal of Advanced Computer Science and Applications Vol.7 No.7, 2016.

[15] Chang, D.Y., "Applications of the extent analysis method on fuzzy AHP," European Journal od Operational Research 95, pp. 649-655, 1996.

[16] Saaty, T.L., "How to make a decision: The Analytic Hierarchy Process," European Journal of Operational Research 48, pp. 9-26, 1990.

[17] Davison, R.M., "A Survey of Group Support Systems: Technology and Operation," City University of Hong Kong, China. Sprouts: Working Papers on Information Systems, 1(12), 2001.

Link Breakage Time Prediction Algorithm for Efficient Power and Routing in Unmanned Aerial Vehicle Communication Networks

Haque Nawaz[1]
Dept. Computer Science
SZABIST Karachi
Sindh Madressatul Islam University
Karachi, Sindh, Pakistan

Husnain Mansoor Ali[2]
Dept. Computer Science
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology
Karachi, Sindh, Pakistan

***Abstract*—UAV Communication Networks (UAVCN) comes under the umbrella of Ad hoc Network technology. It has the critical differences with existing wireless networks, which are high mobility, high speed, dynamic updates, and changes in topology due to high movement, which creates the problem of link breakages and affects the routing performance. This problem degrades the performance of UAVCN in terms; it decreases throughput and minimizes the packet delivery ratio. In this paper, we have tried to overcome this problem by considering the received signal power strength (RSPS). We have proposed an algorithm which uses the received signal power strength and time and calculates the link breakage time prediction by using the interpolation method. We have implemented the proposed technique by modifying the OLSR protocol. The extended protocol termed EPOLSR, which efficiently using the signal power strength and time and increasing the performance of UAVCN. The extended protocol implemented by using a research tool network simulator (v3). The metrics received rate, no of received packets, throughput, and packet delivery ratio (PDR) is considered for evaluation. We have examined the proposed EPOLSR with existing routing protocols. It has been observed that the modified protocol performs better concerning all existing evaluated routing approaches.**

Keywords—UAV; link breakage; algorithm; power; RSPS; routing

I. Introduction

In this study, we have explored the emerging area of adhoc networks, which is known as UAVCN. These nodes can be deployed for a specific operation; either it belongs to the military or civilian mission. [1,2]. These network nodes have high movement. However, it changes the frequent topology; that's why link breakages issue affects the routing [3]. The coverage increased by minimizing the interference of UAVs' communication [4]. The environment and terrain affect the UAVs' communication, hence to overcome these obstacles, a hybrid mechanism of unicast and geocast routing used to know the trajectory and location [5]. The authors optimized the route by using the neural network concepts and implemented the Dynamic Source Routing (DSR) protocol for evaluation and optimization [6]. The researchers proposed the hybrid protocol which helps in link establishment [7]. In this study, experimental work carried out by comparing the adhoc on demand routing (AODV), DSDV, and OLSR routing protocol [8].

In this paper, we have organized the work into sections. The first section represents the introduction, and the second section provides the information of OLSR working mechanism, the third section contains a modification of OLSR Hello and Topology Messages by updating the reserved field, the fourth section indicates the prediction of route failure and Link breakage prediction. The fifth section presents the proposed algorithm. The sixth section highlights the research methodology. The seventh section describes the Results, and in the Eighth section, we have concluded.

II. The OLSR Working Mechanism

It is the proactive routing approach that operates two nodes by sending Hello Messages. It discovers the neighbors and maintains the neighbors' table. It updates the topology status and updates the topology information and maintains the routing table. Also, share the MID messages. The neighbor discovery process is mentioned as:

Fig. 1 shows the neighbor discovery process through Hello Messages. The X node forwards an empty message Hello. The Y node receives the message and stores information X as an asymmetric neighbor because the Y address is not available in Hello message. Y node then forwards Hello Message by asserting X is an asymmetric neighbor as soon as X obtains this message and gets its information address and then declares Y as neighbor symmetric. In last, X node incorporates Y in the Hello, it forwards, and Y recognize X as neighbor symmetric as receiving Hello message.

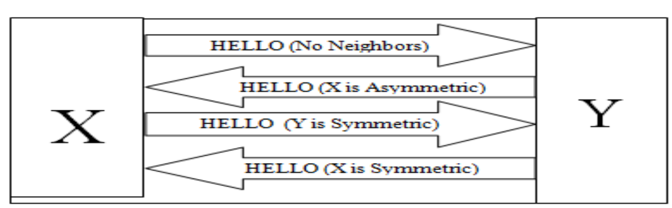


Fig. 1. Neighbor Discovery Process through Hello Messages.

Reserved		Htime	Willingness
Linkcode	Reserved	Link Message service	
Neighbor Interface Address			
Neighbor Interface Address			

Fig. 2. Hello Message Format of OLSR.

The above Fig. 2 shows the Hello Message Format of OLSR. According to RFC 3626, the OLSR furthermore extended or modified by using the reserved field. However, Htime shows the time of the next Hello packet before transmission. The willingness field represents that the willingness of the node when it forwards the traffic. The sender and the neighbor node-link information consists of the link code. It declares the status information of the neighbor node. Link message size provides the total link messages length. The neighbor interface address knows the interface of the neighbor node.

Fig. 3 shows the Topology Control Message Format of OLSR. The Advertised Neighbor Sequence Number (ANSN) shows the incremented sequence number at any time when changes take place in the neighbor set. According to RFC 3626, the OLSR TC packet can furthermore be extended or modified by using the reserved field. The advertised Neighbor Main Address field consists of neighbor node main addresses.

ANSN	Reserved
Advertised Neighbor Main Address	
Advertised Neighbor Main Address	
…………………	

Fig. 3. Topology Control Message Format of OLSR.

III. Modification of OLSR Hello and Topology Messages

The OLSR hello and topology messages are furthermore extended or modified by using the reserved field. The TC packet also extended or modified by using the reserved field. We have updated these packets by adding the power information in terms of signal strength, which is shown in Fig. 4 and Fig. 5. In this study, the modified OLSR termed as EPOLSR. The objective of EPOLSR design was to improve the routing in unmanned aerial vehicle communication networks. The EPOLSR consists of the Routing Table, which stores the routes information of all reachable destination UAVs.

Power		Htime	Willingness
Link code	Reserved	Link Message service	
Neighbor Interface Address			
Neighbor Interface Address			

Fig. 4. Hello Message Format of EP OLSR.

ANSN	Power
Advertised Neighbor Main Address	
Advertised Neighbor Main Address	
…………………	

Fig. 5. Topology Control Message Format of EP OLSR.

When a node receives the Hello message, it performs the following operations.

- To populate one-hop neighbors
- To populate two-hop neighbors
- To perform MPR calculation
- To populate MPR selector set

A. Neighbor Table

The neighbor table contains the following fields.

- Node Address
- Neighbor address identification
- Power Information (LRSPS and CRSPS)
- Time Information (LRSPST and CRSPST)
- Next Link Breakage Time

Each entry in the routing table has the destination address id, next-hop address, data size in packets, delay, and next route breakage time. Based on this information, one hop table and two-hop table neighbor entries, update the routing table. Compare the topology set and update the routing table.

To improve the performance of routing, a method ought to be introduced to administer the route breakages. Although, mostly routing algorithms are used hello messages for the detection of link breakages. The proposed algorithm operates based on link breakage prediction. It estimates link breakage time based on RSPS (received signal power strength) and, it predicts route breakage time as well.

IV. Link Breakage Prediction Time

The hello messages are used to uphold the prediction of link breakage time. Hence, at each time interval, a hello message is broadcasted to everyone instant neighbor of UAV. On account of RSPS intensity, the cost of subsequently, next link breakage time (y) of the analogous neighbor j, might be enlarged or minimized. While a UAV (x) received a hello message from another UAV (y). Then it evaluates the RSPS with the LRSPS from the same UAV. Furthermore, predicting the new cost of NLBT (x,y) using the interpolation method. The given equation (Eq. 1) computes the value of the next link breakage time NLBT(x,y):

$$\mathrm{NLBPT}(x,y) = \mathrm{Crspst} + \frac{(\mathrm{Crspst} - \mathrm{Lrspst}) * (\mathrm{RxThresh} - \mathrm{Crsps})}{(\mathrm{Crsps} - \mathrm{Lrsps})} \quad (1)$$

V. Link Breakage Time Prediction Algorithm for Efficient Power and Routing

Algorithm :

t_1 = Lrspst(last received signal power strength time)
t_2 = Crspst (current received signal power strength time)
pr_1 = *Lrsps*(last received signal power strength)
pr_2 = *Crsps*(current received signal power strength)
$r_x th$ = RxThresh(signal reception threshold)
T_d= Time_d(time difference)
Pkt = pkt(packet)

Begin

Add signal power strength in the reserved field of Hello_Message and send
Hello_Message with this signal power strength to the receiver, it Receive Hello_Message, and observe the received signal power strength, and extract
the signal pwr strength at receiver UAV.
MaxVal = The extracted signal power strength value from HELLO_Message

Input: R_{hp} (received hello pkt)
x (current UAV)
Rspsi received signal power strength intensity

If (R_{hp}.n_ID is not a neighbor of x)
Add_n(R_{hp}.neighbor_ID);
Lrsps = pr_1 (R_{hp}.n_ID);
Lrspst = t_1 (R_{hp}.n_ID);
Update the *Lrsps* of R_{hp}.N_ID;
Update the Lrspst of R_{hp}.N_ID;
If (*Lrsps* ≠ MaxVal)
Begin
Crsps = pr_2 (R_{hp}.n_ID)
Crspst = t_2 (R_{hp}.n_ID)
If (Crspst < Lrspst) // n_UAV move away
Begin
Time_d= (Crspst – Lrspst) * (RxThresh – *Crsps*) / (Crsps – *Lrsps*)
Next Link Breakage Time(R_{hp}*.n_ID)* = Crspst + Time_d
EndIf
Else *Next Link Breakage Time(* R_{hp}*.n_ID)* = MaxVal;
EndIf
End

VI. Research Methodology

The proposed approach or research methodology which we have used is the experimental method. We have used ns-3 (Network Simulator version 3. It can be used for UAVs networks simulation [9, 10]. Using this tool, we have implemented the EP-OLSR routing protocol and compared it with OLSR, DSDV, AODV, and DSR; by using the IEEE 802.11b environment. After the development of the testbed scenario, the simulation was carried out.

VII. Experimental Work and Simulation Parameters

In this scenario, we have used the ns-3 simulation tool and carried out the experiment. The main characteristics of this scenario are shown in Table I.

TABLE. I. CHARACTERISTICS OF THE TESTBED SCENARIO

Parameters	**Values**
Simulation tool	Ns-3.
Adhoc Routing Protocols	EP-OLSR, OLSR, DSDV, AODV, and DSR
Network Scenario Size	1500x400
Number Nodes	25
Data Rate	5 Mbps
Application	Video Streaming
WLAN Physical Characteristics	IEEE 802.11b
Network Protocol	IP
Mobility model	Random Waypoint
Scenario Simulation Time	200 Sec

The scenario has been developed by using 25 nodes — the network scenario based on the area of 1000x400 m2. The WLAN physical characteristics of standard 802.11b are used. The mobility model random waypoint is used. Similarly, the EP OLSR has been developed in C++ and integrated modules in the NS-3 environment. And by using the same wireless LAN physical characteristics standard cross-layer design in the 802.11b environment and the simulation has been run. Similarly, the existing protocols, OLSR, DSDV, AODV, and DSR configured in the scenario, and the simulation has been run. The simulation runs for 200 s. After this experimental setup, the scenario has been accomplished.

VIII. Simulation Results and Discussion

The results show that, in the ns-3 environment, the EP OLSR Receive Rate is 38 Kbps, and on the other hand, we have observed OLSR Receive Rate is 32 Kbps, as depicted in Fig. 6. Overall, the performance of the EP OLSR protocol is found better when compared to that of OLSR.

The results show that, in the ns-3 environment, the EP OLSR Receive Rate is 38 Kbps, and on the other hand, we have observed DSDV Receive Rate is 27 Kbps, as depicted in Fig. 7. Overall, it has been found that the performance of the EP OLSR protocol is better when compared with DSDV.

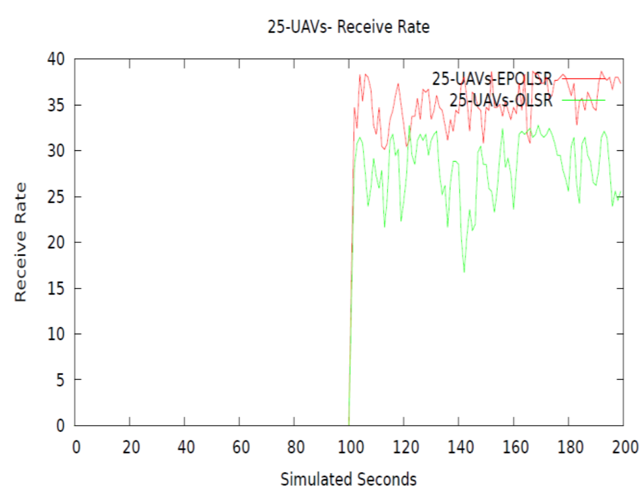


Fig. 6. 25-UAVs- Receive Rate (kbps) using EP OLSR vs. OLSR.

Fig. 7. 25-UAVs- Receive Rate (kbps) using EP OLSR vs. DSDV.

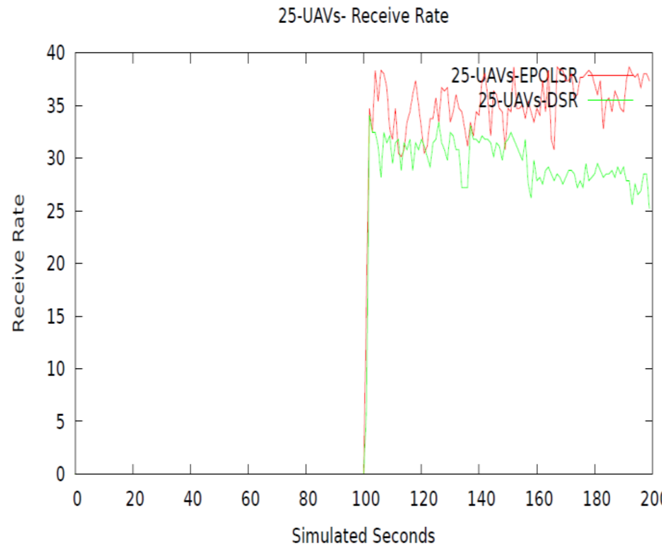


Fig. 9. 5-UAVs- Receive Rate (kbps) using EP OLSR vs. DSR.

The results show that, in the ns-3 environment, the EP OLSR Receive Rate is 38 Kbps, and on the other hand, we have observed AODV Receive Rate is 28 Kbps, as depicted in Fig. 8. Overall, the performance of the EP OLSR protocol is better when compared with AODV.

The results show that, in the ns-3 environment, the EP OLSR Receive Rate is 38 Kbps, and on the other hand, we have observed DSR Receive Rate is 30 Kbps, as depicted in Fig. 9. Overall, the performance of the EP OLSR protocol is found better when compared to DSR.

The results show that, in the ns-3 environment, the EP OLSR Receive Rate is 38 Kbps, OLSR Receive Rate is 32 Kbps, DSDV Receive Rate is 27 Kbps, AODV No of Packets Received is 28 Kbps, on the other hand, DSR Receive Rate is 30 Kbps, as depicted in Fig. 10. Overall, it has been observed that the performance of the EP OLSR protocol is better when compared to OLSR, DSDV, AODV, and DSR.

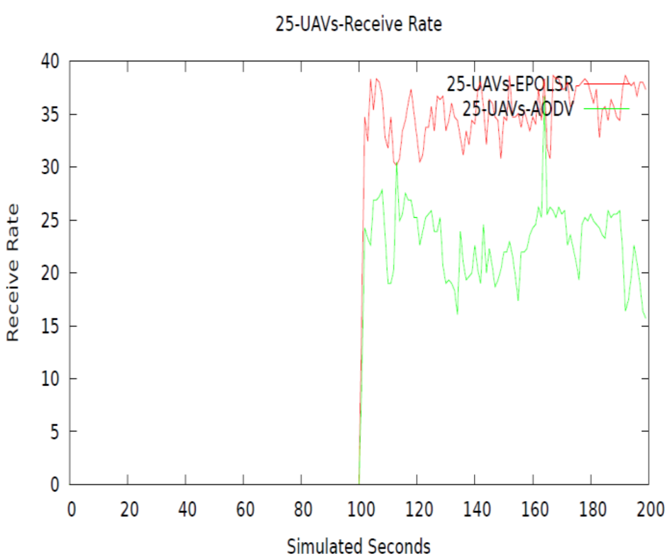


Fig. 8. 25-UAVs- Receive Rate (kbps) using EP OLSR vs. AODV.

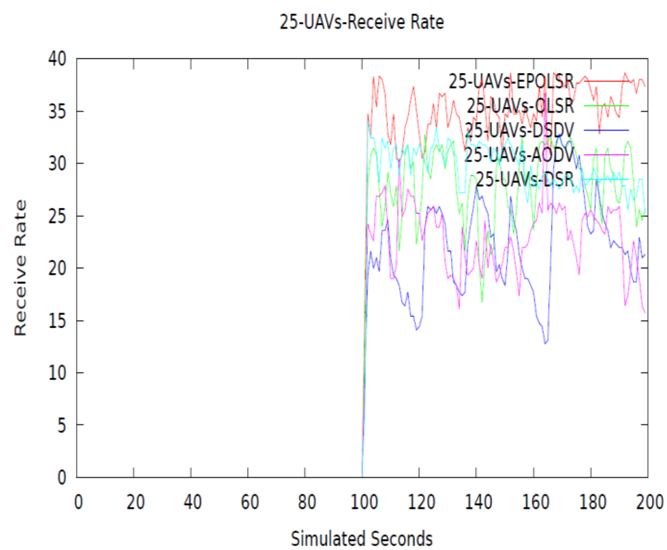


Fig. 10. 25-UAVs- Receive Rate (kbps) using EP OLSR vs. OLSR, DSDV, AODV, and DSR.

The results show that, in the ns-3 environment, the EP OLSR No of Packet Received is 38 Kbps, and on the other hand, we have observed OLSR No of Packet Received is 32 Kbps, as depicted in Fig. 11. Overall, the performance of the EP OLSR protocol is found better when compared with OLSR.

The results show that, in the ns-3 environment, the EP OLSR No of Packet Received is 38 Kbps, and on the other hand, we have observed DSDV No of Packet Received is 26 Kbps, as depicted in Fig. 12. Overall, the performance of the EP OLSR protocol is found better when compared with DSDV.

The results show that, in the ns-3 environment, the EP OLSR No of Packet Received is 38 Kbps, and on the other hand, we have observed AODV No of Packet Received is 26 Kbps, as depicted in Fig. 13. Overall, the performance of the EP OLSR protocol is found better when compared with AODV.

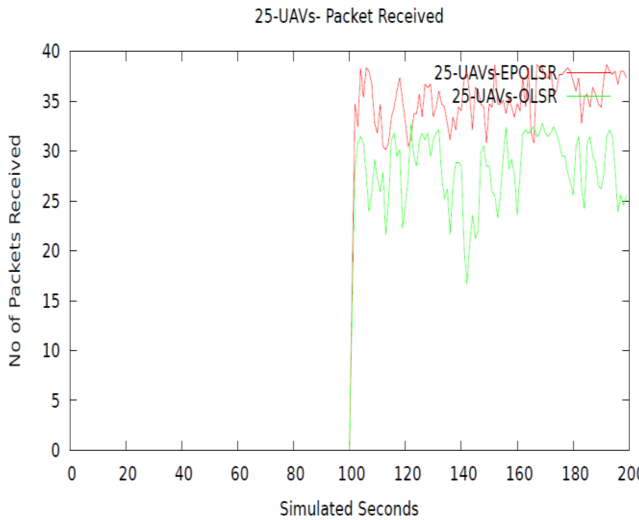


Fig. 11. 25-UAVs- No of Packet Received (kbps) using EP OLSR vs. OLSR.

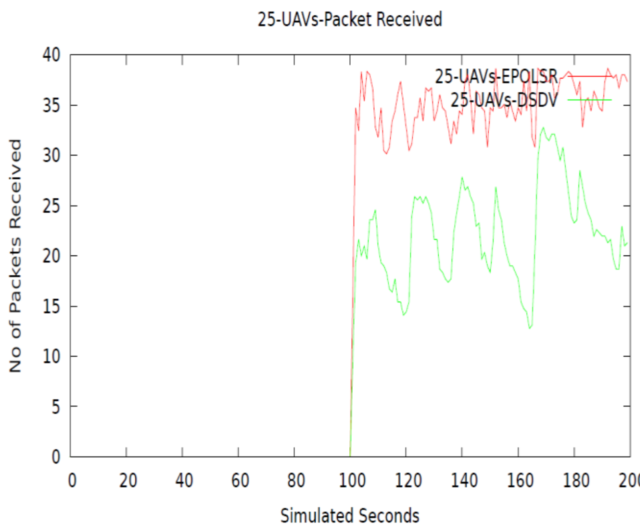


Fig. 12. 25-UAVs- No of Packet Received (kbps) using EP OLSR vs. DSDV.

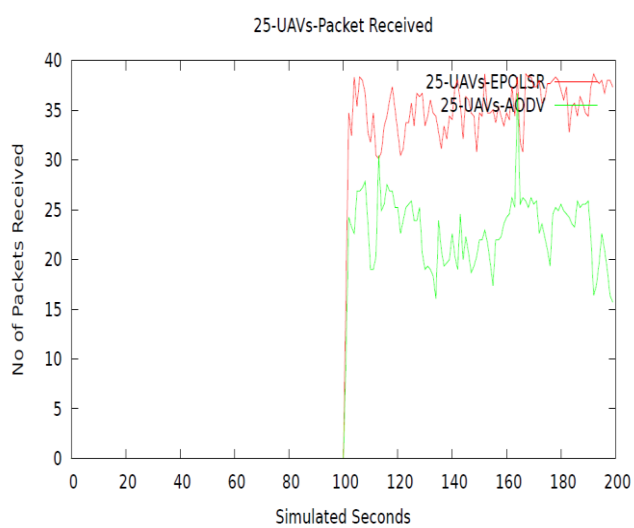


Fig. 13. 25-UAVs- No of Packet Received (kbps) using EP OLSR vs. AODV.

The results show that, in the ns-3 environment, the EP OLSR No of Packet Received is 38 Kbps, and on the other hand, we have observed DSR No of Packet Received is 32 Kbps, as depicted in Fig. 14. Overall, the performance of the EP OLSR protocol is found better when compared with DSR.

The results show that the performance of routing protocols concerning no of the packet received in the ns-3 environment as above depicted Fig. 15. The No of Packets Received in the EP OLSR is 38 Kbps, the No of Packets Received in OLSR is 32 Kbps, the No of Packets Received in DSDV is 25 Kbps, the No of Packets Received in AODV is 26 Kbps, and the No of Packets Received in DSR is 30 Kbps. Overall, it has been observed that the performance of the EP OLSR protocol is better when compared to OLSR, DSDV, AODV, and DSR.

The results show that, in the ns-3 environment, the EP OLSR throughput is 38 Kbps, and on the other hand, we have observed OLSR throughput is 31 Kbps, as depicted in Fig. 16. Overall, the performance of the EP OLSR protocol is found better when compared with OLSR.

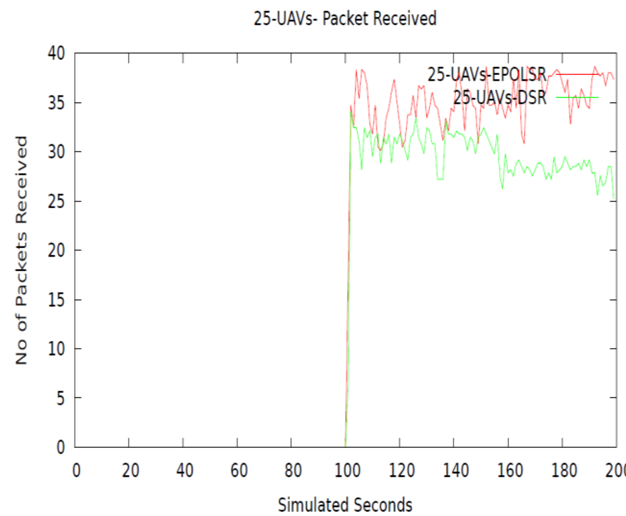


Fig. 14. 25-UAVs- No of Packet Received (kbps) using EP OLSR vs. DSR.

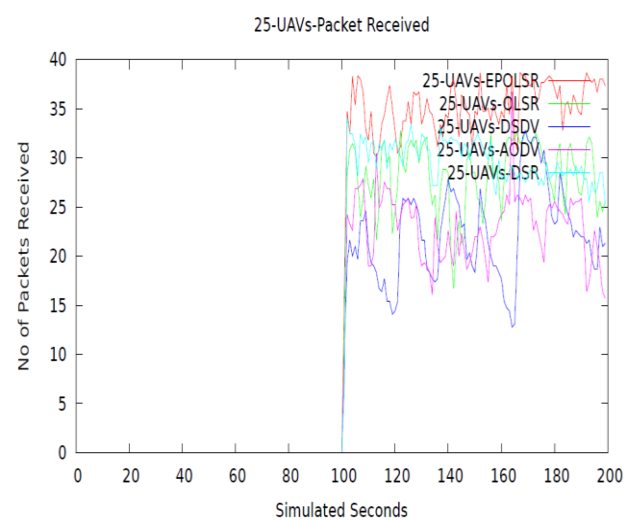


Fig. 15. 25-UAVs- No of Packets Received (kbps) using EP OLSR vs. OLSR, DSDV, AODV, and DSR

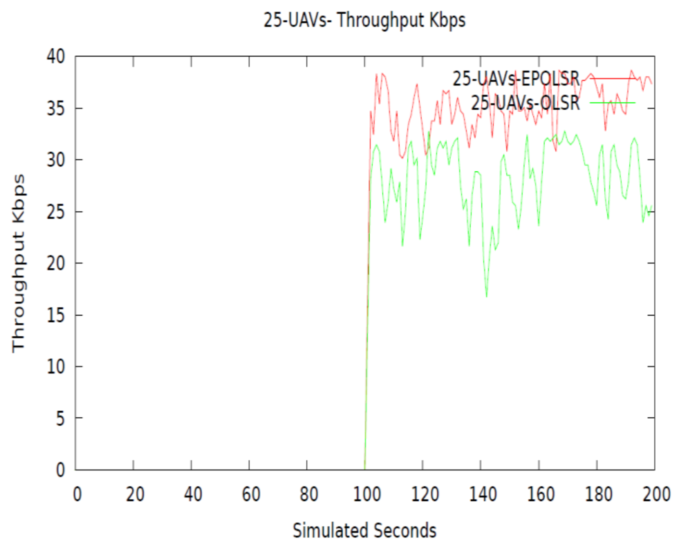


Fig. 16. 25-UAVs- throughput (kbps) using EP OLSR vs. OLSR.

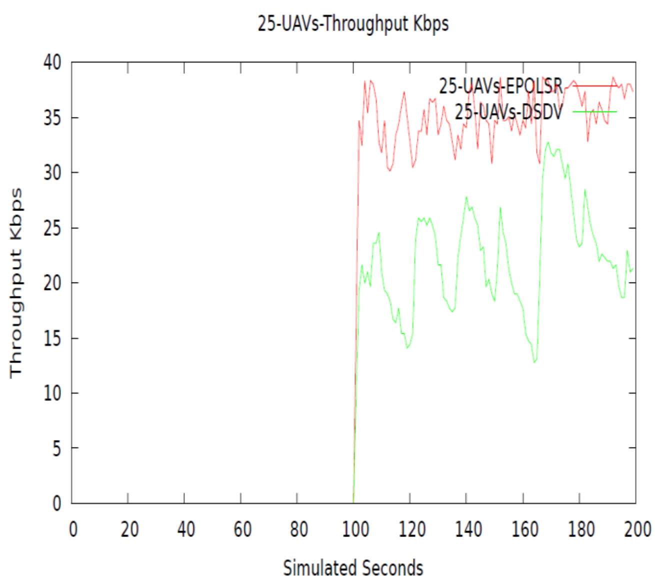


Fig. 17. 25-UAVs- throughput (kbps) using EP OLSR vs. DSDV.

The results show that, in the ns-3 environment, the EP OLSR throughput is 38 Kbps, and on the other hand, we have observed the DSDV throughput is 25 Kbps, as depicted in Fig. 17. Overall, the performance of the EP OLSR protocol is better when compared with DSDV.

The results show that, in the ns-3 environment, the EP OLSR throughput is 38 Kbps, and on the other hand, we have observed AODV throughput is 26 Kbps, as depicted in Fig. 18. Overall, the performance of the EP OLSR protocol is better when compared with AODV.

The results show that, in the ns-3 environment, the EP OLSR throughput is 38 Kbps, and on the other hand, we have observed DSR throughput is 30 Kbps, as depicted in Fig. 19. Overall, the performance of the EP OLSR protocol is better when compared with DSR.

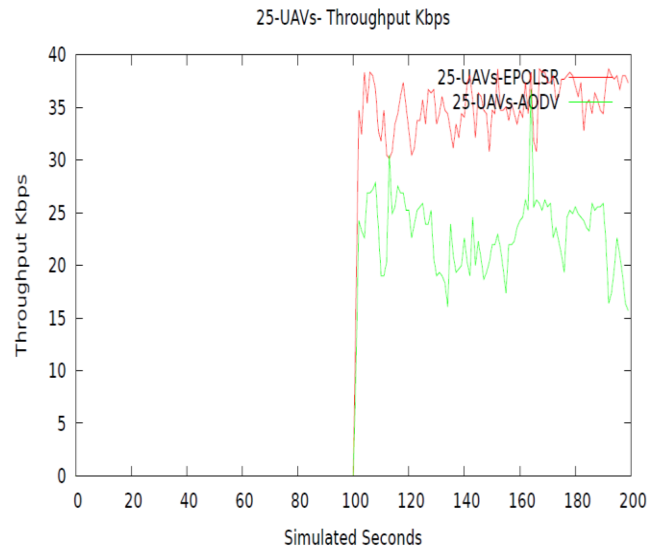


Fig. 18. 25-UAVs- throughput (kbps) using EP OLSR vs. AODV.

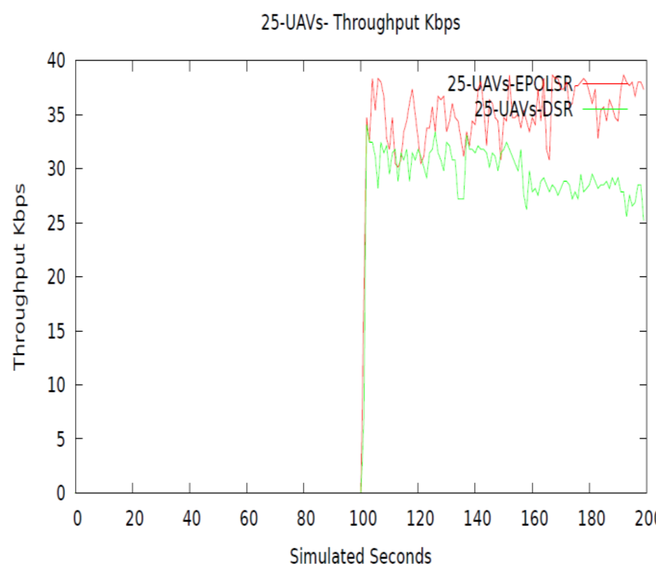


Fig. 19. 25-UAVs- throughput (kbps) using EP OLSR vs. DSR.

The results show that, in the ns-3 environment, the EP OLSR throughput is 38 Kbps, OLSR throughput is 32 Kbps, DSDV throughput is 25 Kbps, AODV throughput is 26 Kbps, on the other hand, DSR throughput is 30 Kbps, as depicted in Fig. 20. Overall, it has been observed that the performance of the EP OLSR protocol is better when compared to OLSR, DSDV, AODV, and DSR.

The results show that, in the ns-3 environment, the EP OLSR Packet Delivery Ratio is 99%, OLSR Packet Delivery Ratio is 96 %, DSDV Packet Delivery Ratio is 83%, AODV Packet Delivery Ratio is 93%, on the other hand, DSR Packet Delivery Ratio is 91 %, as depicted in Fig. 21. Overall, it has been observed that the performance of the EP OLSR protocol is better than OLSR, DSDV, AODV, and DSR.

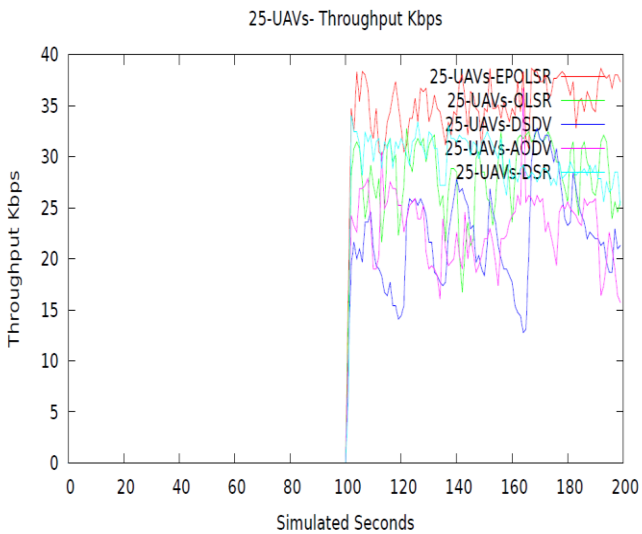


Fig. 20. 25-UAVs- throughput (kbps) *u*sing EP OLSR vs. OLSR, DSDV, AODV, and DSR

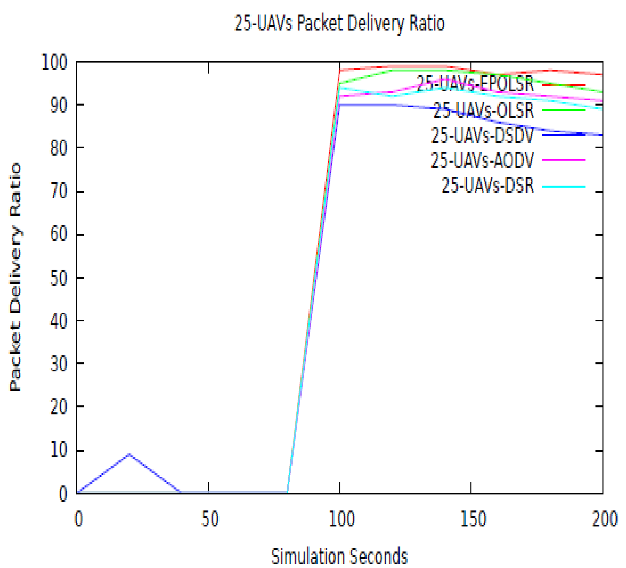


Fig. 21. 25-UAVs- Packet Delivery Ratio *u*sing EP OLSR vs. OLSR, DSDV, AODV, and DSR.

IX. Conclusion

In this research, the UAVCN scenario was developed by using the NS-3. We have simulated 25 UAV nodes and implemented the proposed modified approach of routing, which is termed as EPOLSR. After this, we have evaluated the existing routing protocols with a new approach. In this paper, we have considered and used the received signal power strength (RSPS) by using an algorithm that predicts the link breakage time before route failure. And make the nodes intelligent through hello messages to ensure communication by changing route before failure. However, this method is using the interpolation method and updating the neighbor tables by power information. At the reception node, the signal power strength and time calculated for the next link breakage time by using the interpolation method. We have implemented the proposed technique by modifying the protocol OLSR. The extended protocol termed EPOLSR, which efficiently using the signal power strength and time and increasing the performance of UAVCN. The extended protocol implemented by using the research tool. The metrics received rate, no of received packets, throughput, and packet delivery ratio (PDR) are considered for evaluation. We have examined the proposed EPOLSR with existing routing protocols. It has been observed that the modified protocol performs better concerning all existing evaluated routing approaches.

Acknowledgment

This work has been carried out at SZABIST for the degree of Ph.D. I am very thankful to the administration of SZABIST for providing me the resources to complete this research. Moreover, I am pleased to the reviewers for their comments and grateful to the editor of IJACSA for providing the platform to publish the manuscript.

References

[1] I. Bekmezci, O.K. Sahingoz, and Ş. Temelm, “Flying ad-hoc networks (FANETs): A survey." Ad Hoc Networks, 11(3), pp.1254-1270, 2013.

[2] H. Shakhatreh, A.H. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N.S. Othman, A. Khreishah, and M.Guizani, “Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges." IEEE Access, 7, pp.48572-48634, 2019.

[3] L. Gupta, R. Jain, and G. Vaszkun, “Survey of important issues in UAV communication networks." IEEE Communications Surveys & Tutorials, 18(2), pp.1123-1152, 2015.

[4] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, “Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage." IEEE Communications Letters, 20(8), pp.1647-1650, 2016.

[5] G. Gankhuyag, A.P. Shrestha, and S.J. Yoo, “Robust and reliable predictive routing strategy for flying ad-hoc networks” IEEE Access, 5, pp.643-654, 2017.

[6] H. Yang and Z. Liu, "An optimization routing protocol for FANETs." EURASIP Journal on Wireless Communications and Networking, 2019:120, pp. 1-8, 2019.

[7] Z. Zheng, A.K. Sangaiah, and T. Wang, “Adaptive communication protocols in flying ad hoc network." IEEE Communications Magazine, 56(1), pp.136-142, 2018.

[8] K. Singh, and A.K. Verma, “Experimental analysis of AODV, DSDV and OLSR routing protocol for flying adhoc networks (FANETs)." In IEEE International Conference on Electrical, Computer, and Communication Technologies (ICECCT) (pp. 1-4), 2015.

[9] M.A. Khan, R. Hamila, M.S. Kiranyaz, and M Gabbouj, “A Novel UAV-Aided Network Architecture Using Wi-Fi Direct." IEEE Access, Vol.7, pp. 67305- 67318, 2019.

[10] I. Mahmud, and Y.Z. Cho, “Adaptive Hello Interval in FANET Routing Protocols for Green UAVs” IEEE Access, 7, pp.63004-63015, 2019.

IoT System for Sleep Quality Monitoring using Ballistocardiography Sensor

Nico Surantha[1], C'zuko Adiwiputra[2], Oei Kurniawan Utomo[3], Sani Muhamad Isa[4], Benfano Soewito[5]
Binus Graduate Program – Master of Computer Science
Bina Nusantara University, Jakarta, Indonesia

***Abstract*—Sleep is very important for people to preserve their physical and mental health. The development of the ballistocardiography (BCG) sensor enables the possibility of day-to-day and portable monitoring at home. The goal of this study is to develop an IoT sleep quality monitoring system using BCG sensors, microcontrollers and cloud servers. The BCG sensor produces ECG data from the physical activity of the patient. The data is sent to the sensor and is read by the microcontroller. The sensor data is collected and pre-processed in the microcontroller. The microcontroller then transmits the data obtained from the BCG sensor to the cloud server for further analysis, i.e. to assess the sleep quality. The assessment of data transmission efficiency and resource consumption is carried out in this paper. The findings of the evaluation show that the proposed method achieves higher efficiency, lower response time and decreases memory usage by up to 77% compared to the conventional method.**

Keywords*—*Internet-of-Things; sleep quality; ballisto-cardiography; HRV; ECG

I. Introduction

Sleep is a daily rhythm or circadian rhythm in humans that is governed by the human biological clock in the brain's core hypothalamus [1]. In sleeping conditions, the body will become more relaxed and have minimal movement due to a decrease in attention to the surrounding environment. Sleep is required by humans in order to sustain human health conditions, because it improves the recovery of the exhausted body when it is in the wake, it helps the cognitive function of the brain due to the cleaning of residual waste in the brain [2].

Humans who do not get enough sleep, which is 7-9 hours a day in a certain period of time, result in people experiencing sleep disorders or sleep disorders [3]. Some problems may arise from sleep disorders and disorders such as cardiovascular disease, obesity and diabetes, psychiatric illness, cancer [4] and sleep apnea [5]. Therefore, there is a need for a supporting method or mechanism to track the quality of human sleep to avoid these issues.

One way to monitor the quality of one's sleep is to use the polysomnography (PSG) process. In its calculation, PSG has several criteria, namely the measurement of brain activity, eye movements, breathing and the body. PSG in the medical world is the gold standard for testing the quality of one's sleep, but it has some drawbacks, e.g. the system is only available in every hospital, the procedure is costly, and the sensor is too obstructive in the human body [6].

The use of sensors on the subject will reduce sleep comfort and will certainly have an impact on the quality of sleep measured. The technology developed as the latest state-of - the-art is ballisticocardiography (BCG) sensors. BCG is a heart rate monitor (ECG) and a respiration sensor that works wirelessly, i.e. without interaction with the body of the person being tested [7].

Data from BCG will be used to develop an IoT network consisting of sensors, microcontrollers and web applications on a cloud server. The system is designed to collect patient sleep data by recording directly using a ballistocardiography monitor, the output of which comes in the form of a dataset for heart rate variability (HRV). Those data will then be used to classify sleep cycles in humans, such as waking, light sleep, deep sleep, and rapid eye movement (REM) [8]. Of the 4 sleep stages, the patient's sleep quality will be classified using a combination of the Weight Extreme Learning Machine (WELM) method and the Particle Swarm Optimization (PSO) method used by Utomo et al. (2019) [9] with an accuracy rate of 78.78% in three sleep classes (NREM, REM, awake) and an accuracy rate of 73.09% in four sleep classes.

This scientific study has been carried out in response to previous research which still needs studies in the construction of a portable, high-performance and low-resource-consumption health monitoring system to monitor sleep quality using BCG sensors. The result of this research is expected to provide options for monitoring the quality of human sleep and to provide reports to medical experts on a periodic basis by presenting the results of the extraction of health data from the implementation of IoT. The cost of medical service is therefore cheaper and the patient does not need to see a doctor every time to check the quality of his sleep [8]. In this paper we evaluate the performance of proposed systems in terms of of data transmission and resource consumption aspects.

The remainder of the paper is structured as follows. The background material is explained in Section II of this document. The related works are listed in Section III of this text. The system design and simulation results are explained in Section IV and Section V, respectively. The conclusion is explained in Section VI.

II. Background Material

Some of the background materials as shown in Fig. 1 are discussed in this section.

This research is supported by Directorate General of Strengthening for Research and Development, Ministry of Research, Technology, and Higher Education, Republic of Indonesia.

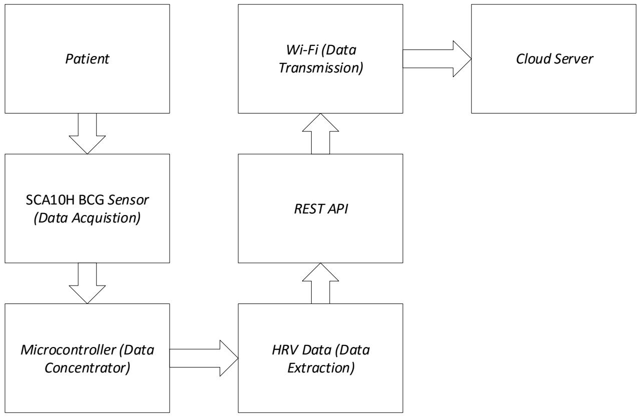


Fig. 1. Background material.

A. SCA11H BCG Sensor

SCA11H is a contactless sleep sensor that uses BCG measurements to monitor patients in the hospital and at home. This system allows users to track patients when they are in bed for sleeping conditions [10]. The power needed by SCA10H is 8-10V, 8mA.

B. Data Acquisition

Data Acquisition in this IoT system is performed by measuring physical data in humans, such as ECG, respiration rate, heart activity, and body posture through sensors. The sensor sends data to the cloud server over the network through a data concentrator [11]. The data concentrator can be a smartphone or a microcontroller.

C. Raspberry Pi

Raspberry Pi is a credit card-sized computer with a weight of only 50 grams that can be purchased at a low price compared to the actual computer. Various devices such as mouse, keyboard, Wi-Fi adapter can be connected to four accessible USB 2.0 ports. Raspberry Pi 3 model B + uses a power of 5v, 700mA [12].

D. Data Concentrator

Data Concentrator works to collect data in real time and to send data to data storage in real time, such as a cloud database. Data Concentrator controls communication protocols for sensor data transmission over a certain period of time and interval [13].

E. Data Transmission

Data Transmission is responsible for delivering data from the data concentrator to a data center or cloud device with assured security and privacy. Data transmission is required because, generally speaking, the data acquisition device on the sensor is only equipped with short-range radio technology such as Bluetooth. Data can be sent to long-term data storage via an internet connection in a data concentrater, e.g. via a Wi-Fi network [14].

F. REST

Representational State Transfer (REST) is an architectural design of the web where the architectural strength in the form of a series of constraints is applied to every element contained in the architecture. This makes it easier for a web designer to design a system and impose behavior on the system to be built [15].

III. Related Works

Several methods have been developed to detect portable and wireless heart signals (ECG). Nandakumar et al. (2015) [16] proposes the detection of sleep apnea based on chest and abdomen movements using a smartphone modified to sonar. Smartphone speakers are set to emit waves with a certain frequency and the reflection of the waves is captured by the microphone on the smartphone. The problem is with the non-standard speakers and microphones on each smartphone, and the system is vulnerable to the sound / movement of other body parts besides the chest and abdomen.

The EarlySense sensor proposed by Tal et al. (2017) [17] as a contactless sensor solution for detecting ECG signals The EarlySense sensor must be placed under the bed and will monitor ECG signals while we are sleeping. EarlySense sensor has a problem that is the placement of sensors must be under the bed which will block the bed which will result in sleep becoming uncomfortable. Sensors are also prone to shifting positions that will impact ECG readings.

Araujo et al. (2018) [18] proposes ApneaLink, which is a portable PSG sensor that is used to read data while the user is sleeping and processes the raw data obtained into a sleep apnea analysis. To be able to do the reading, the ApneaLink sensor is bound to the chest and user's stomach. This will reduce the user's comfort while sleeping.

Yacchirema et al. (2018) [19] have developed an IoT device to monitor the quality of health, specifically related to sleep apnea problems. There are three layers proposed by Yacchirema et al. (2018). They are the IoT layer consisting of sensors, the fog computing layer consisting of a data aggregator, and the cloud layer consisting of a database and user interface. The technology used is HDFS, Apache Spark, Kafka, and Cassandra.

Utomo et al. (2019) have carried out research on the IoT platform to analyze the sleep classification or stages of sleep [9]. The ECG signals are selected as data (input) because the ECG sensor is more simple to use. Utomo et al. (2019) have compared several methods for analyzing sleep cycles, including Extreme Learning Machine (ELM), Backpropagation Neural Network (BPNN), and Support Vector Machine (SVM) designed by Lesmana et al. (2018) [20] and proposed to used Weight Extreme Learning Machine (WELM) with combination with Particle Swarm Optimization to counter the imbalance dataset between sleep classes in sleep sensor.

The SCA11H sensor is a BCG sensor developed by the Murata Electronics company [21]. SCA11H is a BCG (contactless) sensor that can be used to measure ECG signals, which has a correlation with PSG for heart rate signal measurements of 0.97 and has an average error of -0.1 ± 4.4 beats per minute for a 95% confidence level [22]. SCA11H sensor has the advantage that it can be installed at the bedside and read data can be sent via a microcontroller connected to a Wi-Fi module.

From previous research, there is a need for portable and regular sleep monitoring system. Many sensor technologies have been explored for this purpose, e.g. mobile phone, wearable sensor, and portable ECG sensor. The emergence of BCG sensor can introduce the possibility of portable and daily monitoring at home. Because BCG is mounted on the bedside, it is supposed to be functional and comfortable for the patient. Therefore, the sleep monitoring system based on BCG sensor is proposed in this research.

IV. System Design

The main objective of the system design is to introduce an IoT system that focuses only on data acquisition, data concentration and data transmission. Development in cloud server and monitoring application has already been done by Utomo et al. (2019) [23]. This system is being reused in this research with some rework and simplification of some services to increase the performance of resources such as memory, CPU, and disk used on a cloud server. The complete architecture of the proposed system is shown in Fig. 2.

A. Data Acquisition

The data acquisition section consists of a SCA11H sensor which is a contactless BCG sensor. The SCA11H sensor has the ability to record heart and respiratory signals. Data read by the sensor will be sent using Wi-Fi (IEEE 802.11). The SCA11H sensor requires power of 9V-8mA and is installed at the bedside to operate. This data acquisition is expected to produce heart rate data to be sent to the data concentrator.

BCG sensor generates data stored on the microcontroller in the form of a text file printed by the Python program and stored also on a cloud server database using the PostgreSQL database. BCG SCA11h sensor variable output is ordered as follow: timestamp, HR, RR, SV, HRV, SS, status, B2B,

B2B1, B2B2. The explanation of each output is explained in Table 1.

Table 1 shows that the data obtained from SCA11H has a lot of payload output with each type and unit. Data is written in one text file, separated by commas between variables for output data between variables and separated by enter to separate data every second.

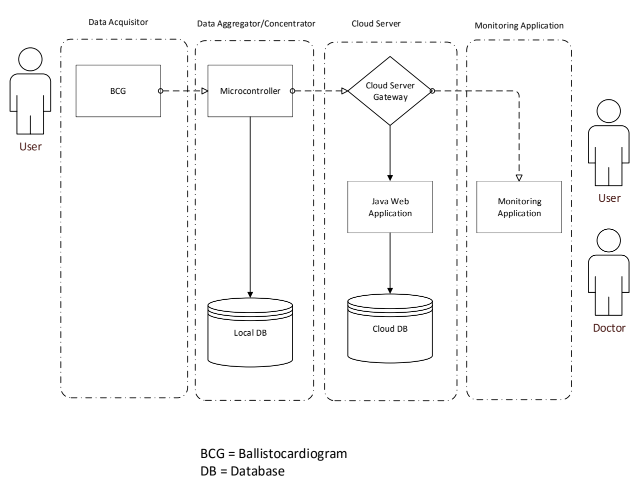


Fig. 2. System Design.

TABLE. I. BCG Sensor Output Data

Variable	Data Type	Unit	Description
time_stamp	S32		System Timestamp
HR	S32	1/min	Heart Rate
RR	S32	1/min	Respiration Rate
SV	S32	ml	Relative Stroke Volume
HRV	S32	ms	Heart rate variability
signal_strength	S32	arbitary unit	measured signal strength indication
status	S32	ms	0 = low signal 1 = ok signal 2 = high signal 3 = (close to) overload 4 = (close to) max HR
B2B	S32	ms	Beat-to-beat time
B2B1	S32	ms	Beat-to-beat time
B2B2	S32	ms	Beat-to-beat time

BCG SCA11H is configured using 1kHz acceleration intervals and runs an algorithm designed by Murata Electronic. Data from the results of the preprocessing algorithm are sent in the output data @ 1Hz rate, which is one data per second. Data at a rate of 1 Hz is also applied when the data representation is issued in the form of a log file in text format.

The study only uses data B2B, B2B1, B2B2 with integer data type type S32. Millisecond value in the process of the unit will be changed to second. This change was made because the preprocessing data made on the cloud server only needs to use B2B data in second units as a reference to classify sleep quality.

B. Data Concentrator

Data concentrator is implemented on Raspberry Pi 3 model B + board. Data from the SCA11H sensor will be sent to the data concentrator, in the form of an endpoint that consists of:

1) Local gateway, which has the function to process raw data storage into a database / database. In addition, the local gateway also functions to order the cloud to perform classification stages of sleep operations.

2) Database to store raw data in the form of text, which has the function to store data from the BCG sensor.

In the data concentrator, there are three pre-process that are applied before transmitting data to the cloud:

1) Data cleansing: is carried out in the data extraction process due to irrelevant ECG data records. To find out the occupancy of the patient's bed, this study discarded the status variable 0 where the status 0 means that the patient's heart rate was not found on the bed that has been installed with BCG SCA11H sensor.

2) B2B data extraction: From all the sensor output as defined by Table 1, only the last 3 output are utilized in this research, i.e. B2B, B2B1, B2B2.

3) B2B data filtering: At this stage if the data obtained has B2B data > 3000 milliseconds, then the data can be removed because this is noise [22]. Furthermore, it also removes unused beat-to-beat empty slots, *i.e.* beat-to-beat which has a value of 0 based on a separate comma that separates data reading. The units in B2B are still in the form of milliseconds, so they need to be changed into second before being sent to the cloud server. The number of B2B samples before and after filtering is shown by Table 2.

TABLE. II. SAMPLE OF BEAT-TO-BEAT (B2B) DATA

Data	Number of B2B sample (previous)	Number of removed B2B sample	Number of B2B sample (after removal)
DemoData1.txt	6741	4034	2707
DemoData2.txt	25602	17413	8189
DemoData3.txt	70224	55159	15065

C. Cloud Server

The cloud server section has a function to collect daily sleep quality data from all existing patients as a doctor's reference. On the cloud server the classification process of sleep stages and the level of sleep quality will be carried out. Cloud server will send data that will be displayed on the monitoring application dashboard.

D. Monitoring Application

The outcomes of patient sleep monitoring will be shown in a tracking report that can be accessed by patients and health professionals through a web application that can be accessed via a smartphone or a laptop. Information to be obtained by patients and health practitioners at this level may be in the form of tables and graphs that provide information on the quality of sleep monitoring of patients in a specific time period. In this study, the monitoring application is developed using Java Application using the JHipster framework where there is an HTML, CSS, and AngularJS framework as a front-end framework and Java version 1.8.0 181 is used as a back-end application support.

E. System Integration

The BCG sensor transmits data on the heart rate to the raspberry pi, which serves as a data focus. The system aggregator / concentrator data includes a program that consists of a python library and a source code. The data concentrator is capable of reading data from the sensor every second with this device. The Python source code named SleepMonitoring. Py will run automatically every time the Raspberry Pi hardware is switched on. The python program can perform pre-processing of heart rate data in the form of data selection, data extraction, and data cleansing.

Data that has been pre-processed will be sent directly to the web application hosted on the cloud server using the RESTful Web Services post form. Patient sleep data will be shown in real-time in the monitoring program. If the patient has finished sleeping, the program on the web server will conduct a quantification process to determine the patient's sleep and sleep quality.

This study uses a service that was developed using Python programming in Raspberry Pi. For all hardware and software used are:

- Microcontroller: Raspberry Pi 3 model b +
- Python version 3
- Raspbian OS version 4.19
- Google Cloud Platform
- SCA10H sensor
- USB to serial TTL Converter PL 2303 YP-01
- Battery 9v as a power supply for the SCA10H sensor
- Charger Adapter for Raspberry Pi
- 2.4GHz and 5 GHz 802.11 ac Wireless LAN on Raspberry Pi 3 model 3+ as a raspberry pi connection to the internet.

V. RESULTS AND DISCUSSION

In this study, we conducted an examination of 26-year-old patients with BCG sensors for three consecutive days that obtained HRV data generating data with the names DemoData1.txt, DemoData2.txt, and DemoData3.txt. This data is stored on the microcontroller in the form of a text file printed by the Python program and stored also on a cloud server database using the PostgreSQL database. In this section, two kinds of evaluation are performed, i.e. data transmission evaluation and resources consumption.

A. Data Transmission Evaluation

Previous study, Utomo et al. (2019) [23] uses simulation data in the form of S32 integer HRV data with the second unit value. It also uses the timestamp in the request body for the REST payload post. However, it is not reliable due to the delay when data transmission is performed. Delays in the REST Architecture cause HRV data out of sync when compared to the input time in the simulator data with the data receive time on the cloud server. Therefore, in this research, we use time_stamp and HRV data directly obtained from the SCA10H sensor in real time.

Data transmission and response time testing samples are obtained from Table 1 where the final B2B sample number is sent every 1 second. The error rate parameter in Table 3 is determined by where the error rate is 0%. This result means that all responses received by the microcontroller from the cloud server are HTTP Status 200 OK. In Table 3, performance measurements are carried out to determine the performance of data transmission (average response time, average sent bandwidth, and average received bandwidth). In this evaluation also obtained a delay on the REST architecture is approximately 9-10 ms.

In order to evaluate the amount of transactions that can be received on a cloud server in one second, then testing is carried out to send 1000 data with ramp-up 0 seconds. The data tested are DemoData1.txt, DemoData2.txt, DemoData3.txt. Each data will be tested 20 times to get the average throughput for this implementation. The result of throughput is shown by Fig. 3.

The average of throughput for proposed method is approximately 913 transactions/second which is higher than the conventional method [23]. While, the response time of the proposed method is 224.7 ms which is lower than the conventional method [23]. The comparison of throughput and response time is shown by Table 4.

B. Resources Consumption Evaluation

In previous study, the monolith applications designed by Utomo et al. (2019) [23] consumes very large memory that is equal to 1,437.67MB for the use of 1 instance server. Therefore, this study tries to reuse previous research and simplify services built on the Java Application that was built on previous research.

To reduce the burden of performance on the cloud server, this study also modified the Tomcat JDBC database connection pool into the Hikari Connection Pool (HikariCP) database connection pool used in previous study. As shown by Fig. 4, Fig. 5 and Fig. 6, the CPU utilization, storage disk utilization, and RAM consumption are 0.016 mCPU Core, 4.592 MB, and 690.2 MB, respectively. From Table 5, it can also be seen that the proposed method can reduce the memory consumption until 77% from the conventional method [23].

TABLE. III. DATA TRANSMISSION EVALUATION

Parameter	DemoData 1.txt	DemoData 2.txt	DemoData3 .txt
HTTP Request from microcontroller	2707	8189	15065
Evaluation Duration	45 Min 7 Sec	2 H 16 Min 29 Sec	4 Hour 11 Min 5 Sec
Error Rate	0%	0%	0%
Average Response From Cloud	9ms	10ms	9ms
Average Received Bandwidth	0.25KB/sec	0.26KB/sec	0.25KB/sec
Average Sent Bandwidth	0.28KB/sec	0.29KB/sec	0.27KB/sec

TABLE. IV. THROUGHPUT AND RESPONSE TIME COMPARISON

Method	Throughput (transactions/s)	Response Time(ms)
Proposed Method	913	224,7
Conventional Method [23]	803,24	253,34

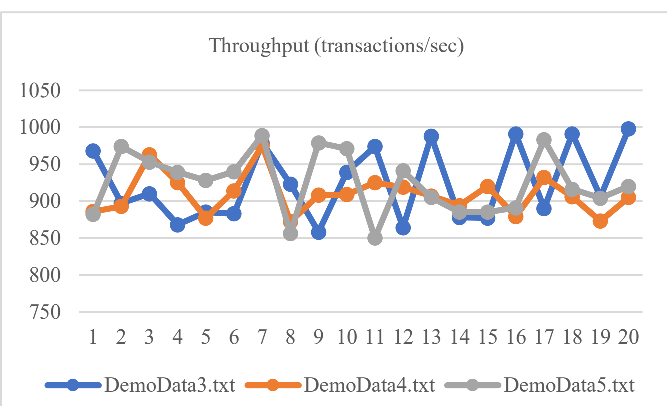


Fig. 3. Throughput Result with RAMP-up 0s.

Fig. 4. CPU Utilization During Monitoring.

Fig. 5. Storage Consumption.

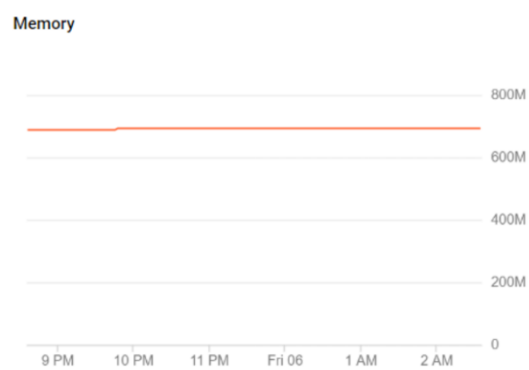


Fig. 6. RAM Consumption.

TABLE. V. RESOURCES CONSUMPTION COMPARISON

Resources Name	Proposed Method	Conventional Method [23]
CPU	0.016 mCPU Core	-
Memory	690.2 MB	2898 MB
Disk	4.592 MB	-

VI. CONCLUSION

In this study, we have proposed the IoT system for sleep quality monitoring using BCG sensor. The utilization of BCG sensor introduces the possibility of daily and portable monitoring of sleep quality at home. We have evaluated the performance of the system in terms of data transmission capacity and resources consumption. The evaluation results that the proposed method can achieve higher throughput, lower response time, and lower memory consumption compared to the conventional method. For the future study, it is recommended to use other extracted sensor data, which includes heart rate, respiration rate, stroke volume, heart rate, variability and beat-to-beat-time as a reference for data to determine a person's sleep quality with more parameters and algorithms complexity.

ACKNOWLEDGMENT

This work is supported by the Directorate General of Strengthening for Research and Development, Ministry of Research, Technology, and Higher Education, Republic of Indonesia as a part of Penelitian Terapan Unggulan Perguruan Tinggi Research Grant to Binus University entitled “Prototipe dan Aplikasi Monitoring Kualitas Tidur Portabel berbasis Teknologi Cloud Computing dan Machine Learning” or “Portable Sleep Quality Monitoring Prototype and Application based on Cloud Computing Technology and Machine Learning” with contract number: 039/VR.RTT/IV/2019 and contract date: 29 April 2019.

REFERENCES

[1] R. E. Mistlberger, B. M. Bergmann, W. Waldenar, and A. Rechtschaffen, “Recovery sleep following sleep deprivation in intact and suprachiasmatic nuclei-lesioned rats,” Sleep, vol. 6, no. 3, pp. 217–233, 1983.

[2] W. H. Moorcroft and P. Belcher, Understanding sleep and dreaming. Springer, 2003.

[3] C. C. Panel et al., “Joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society on the recommended amount of sleep for a healthy adult: methodology and discussion,” Sleep, vol. 38, no. 8, pp. 1161–1183, 2015.

[4] A. D. Laposky, E. Van Cauter, and A V Diez-Roux, “Reducing health disparities: the role of sleep deficiency and sleep disorders,” Sleep Med., vol. 18, pp. 3–6, 2016.

[5] B. Sulistyo, N. Surantha, and S. M. Isa, “Sleep Apnea Identification using HRV Features of ECG Signals,” Int. J. Electr. Comput. Eng., vol. 8, no. 6, 2018.

[6] A. T. M. VAN DE WATER, A. Holmes, and D. A. Hurley, “Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography--a systematic review,” J. Sleep Res., vol. 20, no. 1pt2, pp. 183–200, 2011.

[7] O. T. Inan et al., “Ballistocardiography and seismocardiography: a review of recent advances.,” IEEE J. Biomed. Heal. Informatics, vol. 19, no. 4, pp. 1414–1427, 2015.

[8] N. Surantha, G. P. Kusuma, and S. M. Isa, “Internet of things for sleep quality monitoring system: A survey,” in Proceedings - 11th 2016 International Conference on Knowledge, Information and Creativity Support Systems, KICSS 2016, 2017.

[9] O. K. Utomo, N. Surantha, S. M. Isa, and B. Soewito, “Automatic Sleep Stage Classification using Weighted ELM and PSO on Imbalanced Data from Single Lead ECG,” Procedia Comput. Sci., vol. 157, pp. 321–328, 2019.

[10] M. Takano and A. Ueno, “Noncontact In-Bed Measurements of Physiological and Behavioral Signals Using an Integrated Fabric-Sheet Sensing Scheme,” IEEE J. Biomed. Heal. Informatics, vol. 23, no. 2, pp. 618–630, 2019.

[11] M. Hassanalieragh et al., “Health monitoring and management using Internet-of-Things (IoT) sensing with cloud-based processing: Opportunities and challenges,” in Services Computing (SCC), 2015 IEEE International Conference on, 2015, pp. 285–292.

[12] M. S. D. Gupta, V. Patchava, and V. Menezes, “Healthcare based on iot using raspberry pi,” in 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), 2015, pp. 796–799.

[13] H. Gabbar, Smart Energy Grid Engineering. Academic Press, 2016.

[14] P. Tomar and G. Kaur, Examining cloud computing technologies through the internet of things. IGI Global, 2017.

[15] M. Hemdi and R. Deters, “Using REST based protocol to enable ABAC within IoT systems,” in 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016, pp. 1–7.

[16] R. Nandakumar, S. Gollakota, and N. Watson, “Contactless Sleep Apnea Detection on Smartphones,” in Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys ’15, 2015.

[17] A. Tal, Z. Shinar, D. Shaki, S. Codish, and A. Goldbart, “Validation of contact-free sleep monitoring device with comparison to polysomnography,” J. Clin. Sleep Med., 2017.

[18] I. Araújo et al., “Diagnosis of sleep apnea in patients with stable chronic heart failure using a portable sleep test diagnostic device,” Sleep Breath., pp. 1–7, 2018.

[19] D. C. Yacchirema, D. Sarabia-Jacome, C. E. Palau, and M. Esteve, “A Smart System for Sleep Monitoring by Integrating IoT With Big Data Analytics,” IEEE Access, vol. 6, pp. 35988–36001, 2018.

[20] T. Fennia Lesmana, S. Muhamad Isa, and N. Surantha, “Sleep Stage Identification using The Combination of ELM and PSO based on ECG Signal and HRV,” 2018.

[21] U. Meriheinä, “Monitoring of sleep phenomena.” Google Patents, 2017.

[22] S. Nurmi, T. Saaresranta, T. Koivisto, U. Meriheinä, and L. Palva, “Validation of an Accelerometer Based BCG Method for Sleep Analysis,” Aalto Univ. Publ. Ser. Sci. + Technol., 2016.

[23] O. K. Utomo, “IoT Platform Architecture for Sleep Quality Monitoring,” Bina Nusantara University, 2019.

A Critical Review on Adverse Effects of Concept Drift over Machine Learning Classification Models

Syed Muslim Jameel[1]
Department of Computer and Information Sciences
Universiti Teknologi PETRONAS (UTP), Perak
Malaysia

Manzoor Ahmed Hashmani[2], Hitham Alhussain[3]
Department of Computer and Information Sciences,
Center for Research in Data Science (CERDAS), UTP
High-Performance Cloud Computing Centre (HPC3), UTP
Universiti Teknologi PETRONAS (UTP), Perak, Malaysia

Mobashar Rehman[4]
Faculty of Information and Communication Technology
Universiti Tunku Abdul Rahman
Perak, Malaysia

Arif Budiman[5]
Faculty of Computer Science
University of Indonesia
Depok, West Java
Indonesia

***Abstract*—Big Data (BD) is participating in the current computing revolution in a big way. Industries and organizations are utilizing their insights for Business Intelligence using Machine Learning Models (ML-Models). Deep Learning Models (DL-Models) have been proven to be a better selection than Shallow Learning Models (SL-Models). However, the dynamic characteristics of BD introduce many critical issues for DL-Models, Concept Drift (CD) is one of them. CD issue frequently appears in Online Supervised Learning environments in which data trends change over time. The problem may even worsen in the BD environment due to veracity and variability factors. Due to the CD issue, the accuracy of classification results degrades in ML-Models, which may make ML-Models not applicable. Therefore, ML-Models need to adapt quickly to changes to maintain the accuracy level of the results. In current solutions, a substantial improvement in accuracy and adaptability is needed to make ML-Models robust in a non-stationary environment. In the existing literature, the consolidated information on this issue is not available. Therefore, in this study, we have carried out a systematic critical literature review to discuss the Concept Drift taxonomy and identify the adverse effects and existing approaches to mitigate CD.**

Keywords*—Big data classification; machine learning; online supervised learning; concept drift; Adaptive Convolutional Neural Network Extreme Learning Machine (ACNNELM); Meta-Cognitive Online Sequential Extreme Learning Machine (MOSELM); Online Sequential Extreme Learning Machine (OSELM); Real Drift (RD); Virtual Drift (VD); Hybrid Drift (HD); Deep Learning (DL); Shallow Learning (SL); Concept Drift (CD)

I. Introduction

State of the art Big Data (BD) and Machine Learning (ML) is one of the fundamental pillars of the 4th Industrial Revolution (IR 4.0). BD generates from a variety of sources, including scientific research, finance, government, internet search, sensors, documents, image, audio and video, and others. The nature of BD is very complex due to its non-stationary characteristics (volume, velocity, variety, veracity and variability). ML approaches (specifically Deep Learning) are considered the main drivers to utilize BD for intelligence for offline learning scenarios only. However, these approaches failed to maintain performance accuracy during online learning scenarios. One of the online learning scenarios where ML models degrade their performance accuracy due to the non-stationary environment is Concept Drift (CD)[1].

A. Taxonomy of Concept Drift

Dynamic assumptions of data (features of data changes over time) called Concept Drift [2]. The Concept Drift term in Machine Learning (ML) is being recognized as the most critical problem since many decades for traditional data and big data. Many assumptions in ML is by using static data [3]. However, this issue frequently occurs in an Online Machine Learning scenario where these dynamic conditions change frequently. Therefore, due to the addition of new features in data, ML models degrade their performance accuracy or could fail to classify or predict to correct output.

Notably, in Supervised Online ML, the model is learned through the input and output features from data of one-time span and will be likely to predict or classify the output (class category) from another time. The change in features (among both time) is due to various conditions. It could be due to the data format (variety), distribution (variability), or sources (complexity), which change over time. Another term for Concept Drift refers to the classification boundary or clustering centers that continuously change with time elapsing [04]. These conditions will adversely affect the classification performance of the model. In studies the term CD is modeled based on Bayesian decision theory for class output 'c' and input data 'X' as shown in eq (1);

$$P\left(\frac{c}{X}\right) = P(c)\,P\left(\frac{X}{c}\right) / P(X) \tag{1}$$

Where P(c/X), P(c), P(X/c), and P (X) are posterior, prior, conditional, and feature-based probabilities respectively [3]. The possible conditions of Concept Drift arise P(c/X) undergo changes and causes the shift in the class boundary or conditional probabilities (the number of classes increase), this type of Concept Drift is referred to as Real Drift [5].

Furthermore, if the P (X) (feature-wise distribution of data changes) due to insufficient or partial feature representation of existing data distribution (new additional feature adds or some feature updates) called as Virtual Drift [5]. Also, a study introduces Hybrid Drift as a condition P(c/X), and P (X) occurred consequently [2], as shown in Fig. 1. However, few studies discuss possible configuration pattern based on the frequency of drift, gradual drift (when the variety of concepts changes gradually), consecutive drift (when previous concepts reoccur) and sudden drift pattern (when a concept changes/substitutes abruptly) [6], [7], as shown in Fig. 2.

ML models train to classify according to input and output features with a predefined number of classes. If a feature or class-wise distribution changes over time, then ML models will face a substantial degradation in their performance (because ML models do not have prior knowledge of these changes). However, if these ML models retrain according to newly-arrived data, then they are unable to keep knowledge of the recurrent context (previous training knowledge). As shown in Fig. 3.

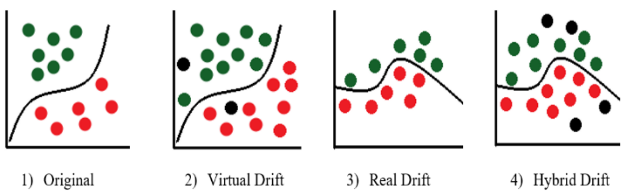


Fig. 1. Types of Concept Drift.

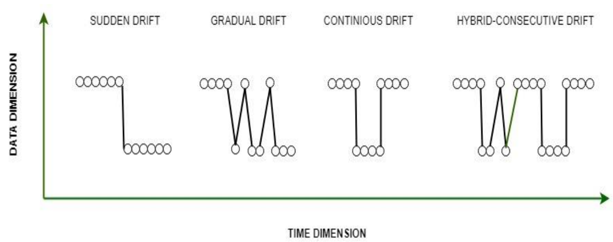


Fig. 2. Configuration Pattern of Concept Drift.

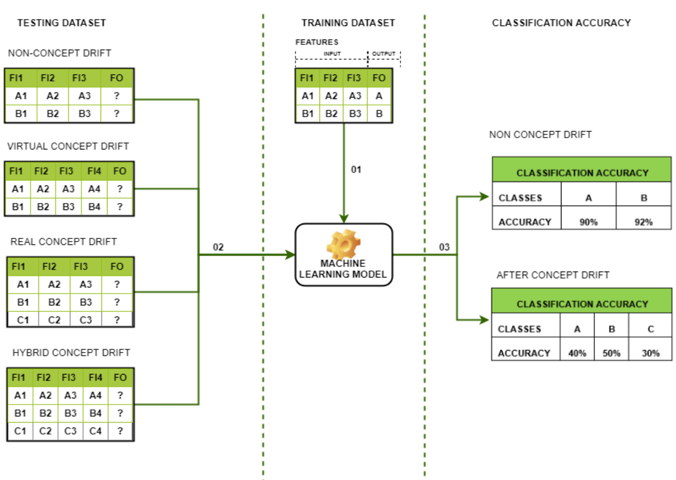


Fig. 3. Performance Degradation in the ML Model Due to Concept Drift.

B. *Causes and Mitigation of Concept Drift in the Classification Problem*

Zilobite I. [8], the classification problem determines through the prior probabilities P(ci) and class conditional probabilities P (x/ci) without considering the Concept Drift scenario. Zilobite I. defines the fixed set of prior probabilities of class and class-conditional as follow;

$$S = \{(P(c1), P\left(\frac{x}{c1}\right)(P(c2), P\left(\frac{x}{c2}\right), \ldots\ldots, (P(ck), P\left(\frac{x}{ck}\right),)\} \quad (2)$$

Where “S” represents the data source at a given time.

Concept Drift based on Bayesian decision theory, as shown below;

$$P(c/X) = P(c)\left\{\frac{P\left(\frac{X}{c}\right)}{P(X)}\right\} \quad (3)$$

$$P(X) = \sum_c P(c)P(X/c) \quad (4)$$

The fundamental causes of the possible change in source data (S) due to P(c/X) are presented by [3][9], which are mentioned below;

P(c) change in class priors (Class Drift)

P(X/c) class distribution permute (Virtual Drift)

P(c/X) posterior distribution of class members changes (Real drift)

P(X/c) + P(c/), class and posterior distribution changes (Hybrid drift)

However, the mitigation strategies are not identical to each type and frequency pattern of Concept Drift (Class, Virtual, Real, continuous, gradual, and sudden or abrupt). For example, we probably like to reuse the past trained classifier if changes reappear (continuous drift pattern), or we may want to suddenly stop classifier and retrain classifier from newly detected changes (abrupt). Thus, to provide a simple approach to handle various types of CD is critical.

In recent studies, researchers propose the term “Adaptability” to avoid performance degradation due to Concept Drift in ML models. The adaptability refers to the feature of ML models (capability) to dynamically adjust itself as per the data changes. This approach allows the ML models to tune or self-regulate for new concept adjustment. Furthermore, this approach possesses the potential to eliminate performance degradation through its dynamic capabilities. However, due to the recurrent context adjustment, the practical implementation for adaptability arises many critical and fundamental questions for researchers. For example, Machine Learning can be categories as context replacement and recurrent context. Context replacement means, how ML models train new concept and forget the previous one. This is simple and can be easily incorporated through the basic adaptability features. Recurrent context refers to how ML models learn a new concept by keeping the previous one. The recurrent context is very challenging. For example, how good accuracy of the ML model can achieve for a new concept and how good it can retain with minimum re-training old data is one of the challenges. In the literature, the adaptability factor can be categorized as semi-adaptive (fundamental dynamic changes in a classifier level) and fully-adaptive (self-regulatory and more autonomous approaches to make classifiers self-regulatory), which are defined in detail in section 02. The fact is that there are few research studies, which provide the practical implementation of recurrent context through the adaptability feature. However, these studies are particular to the type or frequency pattern of Concept Drift or type of data stream. Whereas, the studies define the framework of Concept Drift adaptation in machine learning models present the generalized framework for classifiers. This generalized framework encompasses to make the proper future assumption of data sources, detect all the possible change pattern, tune the classifier parameters or select the appropriate strategy (training, testing or feature manipulation) for specific type of Concept Drift, and optimal model selection (more appropriate model towards the target function) with minimum error rate. Through a dynamic mechanism, in the defined framework, a classifier could be able to regularly evolve and maintain its performance after any Concept Drift., as shown in Fig. 4.

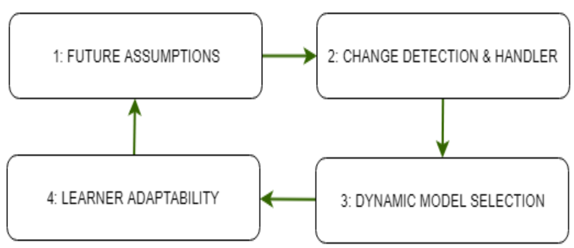


Fig. 4. The Framework for Concept Drift Learners.

II. Approaches to Mitigate the Concept Drift in ML Classification Models

Multidimensional properties of Big Data (BD) and non-stationary nature of Online Supervised Learning (OSL), upfront several new challenges to handle Concept Drift (CD) issue [10]. However, the contribution from the research community to mitigate the adverse effects of CD on Big Data Classification models are rarely reported [2]. Furthermore, most existing approaches are based on Shallow Learning models (ELM, SVM, and others), hence not capable of reasonably handle the non-stationary feature of BD in the OSL scenario [1]. Moreover, several studies urged to adopt these dynamic changes (in classifier) through self-regulatory mechanisms [11] [12] [13].

Shallow Learning approaches (for example, Extreme Learning Machine (ELM), Support Vector Machine (SMV), Multi-Layer Perception Neural Network (MLP NN), Hidden Markov Model, and others.) handle classification and regression problems efficiently in structured data [14][15]. These approaches not perform well for complex unstructured data (Big Data) [1]. However, Deep Learning algorithms found a better selection to handle Big Data stream and extract value with more accuracy over conventional approaches [3]. Besides, some studies ascertained through comprehensive experiments that Deep Learning approaches are appropriate to learn from BD and urged the researcher to explore further new means to handle CD issue due to OSL. A study argued that to find new means to handle Concept Drift in the context of Big Data and OSL is an essential task for the future of Machine Learning [10].

In literature, the issue of CD is mostly handled through the different configurations of the Extreme Learning Machine (ELM). These configurations are either based on a single classifier or ensemble classifier [8][11] [16] [17]. Ensemble classifier considers effective solution than single classifier to improve the classification performance (in terms of accuracy) after CD. Nevertheless, the ensemble approach does not adapt to the numerous drift cases [18] [19], such kind of drift may be handled through the adaptive nature of classifiers.

Few recent studies concentrated towards adaptive learning techniques using ELM based single classifier [2] [15] [19] and ensemble classifier for CD mitigation [20] [21] [22]. However, all these solutions lie in this semi-adaptive category (does not implements the fully autonomous learning behavior). For example, Incremental Data Stream ELM used an incremental approach to train the classifier. In this approach, the number of neuron in hidden layers and selection of the activation layer is dynamic, which enhance the performance of the model. Whereas, this approach handles stream data for gradual drift scenario only [22].

A Dynamic-ELM model uses ELM as a first classifier, whereas the online learning approach is adopted to train the double hidden layer structure of ELM. The improvement in the generalized characteristics of the classifier is incorporated by adding more hidden layers. This approach is capable of mitigating the CD in a short time; however, the performance of this model suffers due to the fast processing speed [21].

Meta-Cognition Online Sequential Extreme Learning Model (MOSELM) proposed for improving class imbalance (binary and multiclass) and Concept Drift for online data classification. This model is first to use Meta-Cognition principles and Online Sequential Extreme Learning Machine (OSELM) but only handle Real Drift [23]. A new adaptive windowing approach is proposed to improve adaptability in Real Drift only [15]. Online Pseudo Inverse Method (OPIUM) is based on Gravel methods, the incremental solutions to computing pseudo-inverse of a matrix. OPIUM tackles the real Concept Drift with the discriminant function boundary shift in streaming data only [19].

A recent study proposed an adaptive ML model (AOSELM) using a single classifier approach based on Online Sequential Extreme Learning Machine (OSELM) [23], and Constructive Sequential Extreme Learning Machine (COSELM) [24] to handle the Concept Drift issue for classification and regression problem. AOSELM is the simple solution used matrix adjustment. Results were satisfactory for handling Real Drift but not satisfactory to handle virtual and Hybrid Drift and did not yield better output on real data. Single classifier results may not exceed the adoptable ensemble or full batch approach due to its shared weight changes [2]. Table I represents the notable contributions (concept drift adaptation models) and highlights its pitfalls.

TABLE. I. LATEST CONCEPT DRIFT HANDLING APPROACHES

Year, Author	**Method**	**Advantages**	**Disadvantages**	**Ref**
1. Dynamic integration of classifiers for handling Concept Drift				
2008, Tsymbal, A	Ensemble Integration	Handle Virtual Drift	No-Adaptability in model Synthetic Dataset used	[16]
2. Reacting to different types of Concept Drift: The accuracy updated ensemble algorithm				
2014, Brzezinski, D	AUE	Better classification accuracy Less memory consumption	Used Synthetic data set Further, need improvement inaccuracy	[25]
3. Reacting to different types of Concept Drift with adaptive and incremental one-class classifiers				
2015, Krawczyk, Bartosz,	AW-SVM classifier	No need to drift detector Provide classification accuracy for stream data.	Further improvement needed in accuracy and adaptability.	[26]
4. Classification of uncertain data streams based on extreme learning machine				
2015, Cao, Keyan, et al	WEC-ELM	Improved efficiency and accuracy during the drift. Deal with real-time uncertain data streams.	Handle only gradual drift Need to work on high dimensional data Not viable for Big Data.	[27]
5. A fast-incremental extreme learning machine algorithm for data streams classification called IDS-ELM				
2016, Xu, Shuliang,	IDS-ELM	Improved performance Suitable for real-time data due to fast processing	Handle gradual drift in streaming Need further improvement	[22]
6. Dynamic extreme learning machine for data stream classification				
2017, Xu, Shuliang	Dynamic-ELM	Improved accuracy Fast processing speed	Accuracy suffers due to speed I need further improvement.	[21]
7. Self-Adaptive Windowing Approach for Handling Complex Concept Drift				
2015, Khamassi, Imen, et al	EDIST2	Concept Drift Detection and handling data streams. Reasonable accuracy rate at Synthetic and Real data Used Self-Adaptive Windowing approach.	Need to handle another type of drift i.e. Virtual.	[28]
8. Meta-Cognitive online sequential extreme learning machine for imbalanced and concept-drifting data classification				
2016, Mirza, Bilal	OS-ELM	Imbalanced and Concept Drift learning model Metacognition is used as self –regulatory	It only works RD drift	[17]
9. Adaptive Online Sequential ELM for Concept Drift Tackling				
2016, Budiman, Arif	AOS-ELM	Handle Real Drift, Virtual Drift, and Hybrid Drift Work well on sudden drift and recurrent Concept Drift type. The public data set is used for regression and classification problem. Good adaptability and retaining recognition	Improvement is needed in VD and HD. Work on specific dataset Need further adaptability factor	[2]
10. Adaptive Convolutional ELM for Concept Drift Handling in Online Stream Data				
2016, Budiman, Arif	CNN+ELM	A hybrid model ACNNELM proposed the integration approach of CNN with ELM. Enhanced CNN features used with ELM. Simple solution good accuracy results and work.	Results non-image data, i.e. voice, action, logging or 1D or more than 4D data. Further improvement needed to increase accuracy. Handling complex variable needed. Need to enhance adaptability factor.	[3]

III. RESULT ANALYSIS AND DEDUCTION

Through the comprehensive literature analysis, we can safely state that the performance degradation in Big Data Classification models (in terms of accuracy) due to Concept Drift is still a critical problem. The existing solutions can be categorized into as follow;

1) Non-adaptive and semi-adaptive (single classifier based) SL approaches.

2) Non-adaptive and semi-adaptive (ensemble classifier based) SL approaches.

3) Non-adaptive and semi-adaptive (single classifier based) DL approaches.

4) Non-adaptive and semi-adaptive (ensemble classifier based) DL approaches.

5) Semi-adaptive Hybrid (DL and SL) approaches.

The existing solutions are either limited to a specific type of CD, or their results are biased towards specific CD conditions or dataset. In addition to that, the classification degradation does not reasonably retain after CD handling for complex datasets (CIFAR 10), as shown in Table II. Table II demonstrates the simulations on the most prominent Big Data classification models under certain CD conditions. The experiments carried out to validate the problem formulation of the Concept Drift issue. In this experiment, we used MNIST [29], Not-MNIST, and CIFAR 10 [30] dataset. The MNIST dataset is recognized as the benchmark dataset for the classification problem, whereas Not-MNIST is an extension of the MNIST dataset, contains some foolish images for providing some challenging data environment. CIFAR 10 is the dataset for color images. MATLAB model R2018a, using Deep Learning toolbox [31] using NVIDIA GeForce GTX 950, 768 GPU cores with 2 GB RAM. The cross-validation and holdout method is used for evaluation and the testing accuracy is measured after a specific type of CD. Interestingly, through the results, we can determine that ACNNELM is better for handling CD for MNIST and Not-MNIST dataset, whereas we found the promising testing accuracy of CNN in CIFAR10 dataset (color images). Also, in our previous study [32], we have performed several experiments to validate the Concept Drift issue.

TABLE. II. LATEST ML MODELS AND THEIR TESTING ACCURACY IN DIFFERENT IMAGE DATASETS

No	Category	ML model	Dataset	Testing Accuracy %
1	Shallow Learning (Non-adaptive)	Support Vector Machine (SVM)	MNIST	83.47
			Not-MNIST	72.65
			CIFAR 10	22.98
2	Deep Learning Model (Non-adaptive)	Convolutional Neural Network (CNN)	MNIST	84.34
			Not-MNIST	72.98
			CIFAR 10	**45.67**
3	Shallow Learning (Semi-adaptive)	Online Sequential ELM (OSELM)	MNIST	82.34
			Not-MNIST	73.98
			CIFAR10	38.68
4	Hybrid Deep Learning (Semi-adaptive)	Adaptive Convolutional ELM (ACNNELM)	**MNIST**	**90.45**
			Not-MNIST	**78.24**
			CIFAR 10	37.56

IV. CONCLUSION

Concept Drift issue can be handled by improving the ML model accuracy and enhancing the adaptability factor. The adaptability feature talks about how an ML model capable of retaining its previous training data knowledge. The ultimate goal of improvement in the ML model and adaptability is handling CD issues, whereas adaptability in the ML model can reduce the computational processing and training time too. According to literature, adaptability can be classified into two types; semi-adaptive (less adaptability) and self –regulatory (a more general aspect of autonomous learning). Current solutions to handle Concept Drift either handle image data or stream data. However, these data classification only provides non-adaptive or semi adaptive solutions (which restrict to utilizing the complete essence of adaptability factor and to handling Concept Drift in Big Data environment). Some possible research directions to overcome CD are; to investigate and formulate the relationship between Concept Drift (Big Data) and exiting Machine Learning models. Quantification and characterization of Concept Drift for Big Data streams. Propose a framework for fully adaptive models for Big Data streams.

Current solutions improved classification accuracy by working on the fewest parameters. For example. The best ML model ACNNELM for Big Data stream classification worked on six parameters which are; Training data composition.

1) Number of kernels
2) Number of layers
3) Type of activation function
4) Number of iteration
5) Variable learning rate

However, to identify the latest critical parameters (for example, hyperspectral features) of advance ML models (Deep Learning) and model the matrix to measure the adaptability factor are potential research directions.

ACKNOWLEDGMENT

This research study is conducted in Universiti Teknologi PETRONAS (UTP), Malaysia as a part of research project "Correlation between Concept Drift Parameters and Performance of Deep Learning Models: Towards Fully Adaptive Deep Learning Models" under Fundamental Research Grant Scheme (FRGS) Ministry of Education (MoE) Malaysia (Grant Reference: FRGS/1/2018/ICT02/UTP/02/2).

REFERENCES

[1] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," J. Big Data, 2015.

[2] A. Budiman, M. I. Fanany, and C. Basaruddin, "Adaptive Online Sequential ELM for Concept Drift Tackling," Comput. Intell. Neurosci., vol. 2016, 2016.

[3] A. Budiman, M. I. Fanany, and C. Basaruddin, "Adaptive Convolutional ELM For Concept Drift Handling in Online Stream Data," Oct. 2016.

[4] W. Zang, P. Zhang, C. Zhou, and L. Guo, "Comparative study between incremental and ensemble learning on data streams: Case study," J. Big Data, vol. 1, no. 1, pp. 1–16, 2014.

[5] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on Concept Drift adaptation," ACM Comput. Surv., vol. 46, no. 4, pp. 1–37, 2014.

[6] L. Kuncheva, Classifier ensembles for changing environments, in: Multiple Classifier Systems, Vol.3077 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2004, pp. 1–15.

[7] L.I.Kuncheva,Classifierensemblesfordetectingconceptchangeinstreaming data: Overview and perspectives, in: 2nd Workshop SUEMA 2008 (ECAI 2008), 2008, pp. 5–10.

[8] Žliobaitė, Indrė. "Learning under concept drift: an overview." arXiv preprint arXiv:1010.4784 (2010).

[9] Kelly, M. G., Hand, D. J., & Adams, N. M. (1999, August). The impact of changing populations on classifier performance. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 367-371). ACM.

[10] M. Rouse., "Predictive Analytics Definition," 2009.

[11] R. Elwell and R. Polikar, "Incremental learning of Concept Drift in nonstationary environments," IEEE Transactions on Neural Networks, vol. 22, no. 10. pp. 1517–1531, 2011.

[12] G. Ditzler and R. Polikar, "Incremental learning of Concept Drift from streaming imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 25, no. 10, pp. 2283–2301, 2013.

[13] I. Zliobaite et al., "Next challenges for adaptive learning systems," ACM SIGKDD Explor. Newsl., vol. 14, no. 1, p. 48, 2012.

[14] G.-B. Huang et al., "Extreme learning machine: Theory and applications," Neurocomputing, vol. 70, no. 1–3, pp. 489–501, 2006.

[15] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," IEEE Trans. Syst. man, Cybern. Part B, Cybern., vol. 42, pp. 513–29.

[16] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen, "Dynamic integration of classifiers for handling Concept Drift," Inf. Fusion, vol. 9, no. 1, pp. 56–68, 2008.

[17] B. Mirza and Z. Lin, "Meta-cognitive online sequential extreme learning machine for imbalanced and concept-drifting data classification," Neural Networks, vol. 80, pp. 79–94, 2016.

[18] N. Liu and H. Wang, "Ensemble based extreme learning machine," IEEE Signal Process. Lett., vol. 17, no. 8, pp. 754–757, 2010.

[19] A. van Schaik and J. Tapson, "Online and adaptive pseudoinverse solutions for ELM weights," Neurocomputing, no. Part A, pp. 233–238, 2015.

[20] J. Zhai, J. Wang, and X. Wang, "Ensemble online sequential extreme learning machine for large dataset classification," in 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2014, pp. 2250–2255.

[21] S. Xu and J. Wang, "Dynamic extreme learning machine for data stream classification," Neurocomputing, vol. 238, pp. 433–449, 2017.

[22] S. Xu and J. Wang, "A fast incremental extreme learning machine algorithm for data streams classification," Expert Systems with Applications, vol. 65. pp. 332–344, 2016.

[23] N.-Y. Liang, G.-B. Huang, P. Saratchandran, and N. Sundararajan, "A Fast and Accurate Online Sequential Learning Algorithm for Feedforward Networks," IEEE Trans. Neural Networks, vol. 17, no. 6, pp. 1411–1423, 2006.

[24] Y. Lan, Y. C. Soh, and G. Bin Huang, "A Constructive Enhancement for Online Sequential Extreme Learning Machine," in Proceedings of the International Joint Conference on Neural Networks, 2009, pp. 1708–1713.

[25] D. Brzezinski and J. Stefanowski, "Reacting to different types of Concept Drift: The accuracy updated ensemble algorithm," IEEE Trans. Neural Networks Learn. Syst., vol. 25, no. 1, pp. 81–94, 2014. [38].

[26] B. Krawczyk and M. Wozniak, "Reacting to different types of Concept Drift with adaptive and incremental one-class classifiers," in Proceedings - 2015 IEEE 2nd International Conference on Cybernetics, CYBCONF 2015, 2015, pp. 30–35.

[27] K. Cao, G. Wang, D. Han, J. Ning, and X. Zhang, "Classification of Uncertain Data Streams Based on Extreme Learning Machine," Cognit. Comput., vol. 7, no. 1, pp. 150–160, Feb. 2015.

[28] I. Khamassi, M. Sayed-Mouchaweh, M. Hammami, and K. GhÃ©dira, "Self-Adaptive Windowing Approach for Handling Complex Concept Drift," Cognit. Comput., vol. 7, no. 6, pp. 772–790, 2015.

[29] http://yann.lecun.com/exdb/mnist/.

[30] Krizhevsky, Alex, Vinod Nair, and Geoffrey Hinton. "The CIFAR-10 dataset." online: http://www. cs. toronto. edu/kriz/cifar. html (2014).

[31] https://www.mathworks.com/help/nnet/ug/deep-learning-in-matlab.html.

[32] Manzoor Ahmed Hashmani, Syed Muslim Jameel, Hitham Al-Hussain, Mobashar Rehman and Arif Budiman, "Accuracy Performance Degradation in Image Classification Models due to Concept Drift" International Journal of Advanced Computer Science and Applications(IJACSA), 10(5), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100552.

Scalability Performance for Low Power Wide Area Network Technology using Multiple Gateways

N.A. Abdul Latiff[1], I.S. Ismail[2]
Faculty of Ocean Engineering Technology and Informatics
University Malaysia Terengganu
Kuala Terengganu, Malaysia

M. H. Yusoff[3]
Faculty of Informatics and Computing
Universiti Sultan Zainal Abidin
Kuala Terengganu, Malaysia

***Abstract*—Low Power Wide Area Network is one of the leading technologies for the Internet of Things. The capability to scale is one of the advantage criteria for a technology to compare to each other. The technology uses a star network topology for communication between the end-node and gateway. The star network topology enables the network to support a large number of end-nodes and with multiple of gateways deployed in the network, it can increase the number of end nodes even more. This paper aims to investigate the performance of the Low Power Wide Area Network Technology, focusing on the capability of the network to scale using multiple gateways as receivers. We model the network system based on the communication behaviours between the end-node and gateways. We also included the communication limit range for the data signal from the end-node to successfully be received by the gateways. The performance of the scalability for the Low Power Wide Area Network Technology is shown by the successfully received packet data at the gateways. The simulation to study the scalability was done based on several parameters, such as the number of end-nodes, gateways, channels and also application time. The results show that the amount of successfully received data signal at gateway increased as the gateways, application time and channel used increased.**

Keywords*—Low power wide area network; scalability; simulation; multiple gateways

I. Introduction

For the past few years, the advances in the development of Internet-of-Things (IoT) speed up the growth for new, innovative, convenience, and economic benefit applications. The things are any devices holding the ability of sensing, computing, and exchanging information with other devices by communicating via the internet [1]. The technology inspired by IoT is believed to be able to enhance the effectiveness of energy consumption, resources management, productivity, and environmental monitoring. The IoT has spread its importance to various domains which previously has a limitation in multiple possible applications, for example, real-time environmental, remote health-care, industrial control, productions systems, smart city, and transportation [2].

Developers have worked on many applications for various uses in the IoT, and specific applications require specialized technology to work accordingly. Traditional short-range communication technologies such as Bluetooth and ZigBee are not ideal for applications that require long-range communication. While cellular technology can provide broader coverage than conventional radio technology, it requires much energy to operate, which is not ideal for low power applications.

Recent advancement in applications that require smaller size device, low power consumption, and cost-effective have shown positive progress in the development of a new communication technology known as the Low Power Wide Area Network (LPWAN). The new technology becomes a complement to conventional communication technology such as cellular and short-range wireless technologies by augmenting a better functionality and requirement for IoT applications. The unique features of LPWAN technology such as high coverage, low bandwidth, and low power consumption, are in line with the requirements of IoT applications that only need to transmit small data sizes remotely.

LPWAN technology is designed to support billions of devices for the various applications of IoT. The technology uses a star topology architecture in which multiple end-nodes communicate directly to the gateway [3]. However, numerous end-nodes transmit the data signal to the gateway and this cause traffic overload, and eventually, there will be data signal loss at the gateway. Increasing the number of gateways can reduce the data signal overload by the single gateway. Even so, the scalability analysis frequently only used a single gateway to study the performance of the LPWAN.

In this study, the performance of the LPWAN in the capability to scale using multiple gateways. The development and simulation of the proposed network model were done based on the collision behaviour of the data signal from the end-node at the gateway using the MATLAB platform. The organization of this paper is as follow: In Section 2, related work on the previous study was presented. Then, the introduction of Low Power Wide Area Networks was presented in Section 3. Next, the proposed network model was discussed in section 4. In Section 5, the simulation procedure and parameters were discussed, and result and discussion were presented in Section 6. Finally, conclude the paper in conclusion.

II. Related Work

The previous work on several studies on model development for LPWAN has been conducted previously for a better understanding of LPWAN's ability to scale. This section presents the previous studies on several works focusing on modelling and scalability of LPWAN.

Georgiou and Raza [3] present the study of modelling of LPWAN to analyse the capability of the technology to scale. The authors used stochastic geometry framework to model the performance of the Long-Range (LoRa) network by using a single gateway. The study proposed an outage probability model which occurs at the gateway called outage condition.

Example of studies following the outage probability model by the authors in [3] was presented in [4]–[6]. The authors in [4] used time diversity to increase the probability of successful packet delivery from the sensor nodes to the gateways. While in [5], the authors presented the paper based on the model in [3]. It included the effect of co-spreading factor (co-SF) interference and inter-SF interference as the model in [3] did not consider the interferences factor in the model. The model presented in [6] is the extending of the outage models from [3] for diversity techniques.

The study of the scalability of LPWAN was presented in [7] by M. C. Bor, U. Roedig, T. Voigt, and J. M. Alonso. This paper investigates the number of transmitters that LoRa network can support. The authors developed LoRa simulator called LoRaSim, which is used to study the scalability of LoRa network. Following the model and simulator from [7], several studies focus on the performance of LPWAN in terms of scalability [8]–[10]. The authors in [8] used LoRa communication model based on [7] to develop further an improved version of LoRaSim called EXPLoRa. Meanwhile, the authors in[10] used the model and simulator from [7] to study the performances of LoRaSim on three different simulation parameters; SF, bandwidth, and coding rate.

The authors in [9] also developed a LoRa model similar to the model from [7]. The scalability of the LoRa network was studied by observing the most significant possible number of LoRa transmitter while satisfying the average packet success probability. The other model was developed by the authors in [11] to study the scalability of LoRa technology. The model used LoRa interference behaviour for the development of the data signal collision model. Meanwhile, the authors in [12] develop LoRaWAN simulator study the scalability of the LPWAN. The development of the packet collision model was inspired by the collision model from [11] to determine the behaviours of the data signal collision and the capture effect. The investigation in [13] shows improvement in the network scalability when using a method which assigning the SF used by the end-nodes in the network.

III. Low Power Wide Area Network

Low Power Wide Area Network is a wireless communication technology that enables end nodes to communicate over long distance using low bit rates and low energy consumption [14][15][16]. Previous studies have shown that LPWAN technology enables the final node to communicate with gates over a distance of 3 kilometres for urban areas, while more than 10 kilometres for rural areas [17]. Additionally, in the line of sight circumstance, the last node data signal can reach a gateway located 20 kilometres away [18] can still reach the gateway as far as 30 kilometres, as reported in [19].

The ability of end-nodes to communicate remotely with a gateway is based on two main special features of the LPWAN, the star network topology and modulation technique. The LPWAN device mostly operates in the unlicensed Industrial, Scientific and Medical (ISM) bands at 169, 433, 868/915 MHz, and 2.4 GHz [20]. However, these frequency values [21][22] depend on the region in which the technology is being used.

Dynamic progress in LPWAN technology development has created many LPWAN-based applications and solutions in the market. The current most known LPWAN technologies are Sigfox and Semtech. The Sigfox technology uses three main components for the communication, which are Ultra Narrow Band radio technology, Binary Phase Shift Keying and Gaussian Frequency Shift Keying modulation. Typically, depending on the region, the ISM band used by the technology is at 868 to 869 MHz and 902 to 928 MHz. Sigfox devices are capable of sending small data with 12 bytes of maximum data size for uplink data while 8 bytes of downlink using the Lightweight protocol. Altogether, the Sigfox frame uses 26 bytes, with 12 bytes of load data and 14 bytes for protocol overhead. This protocol overhead is smaller than conventional LPWAN technology, which applies more significant size protocol overheads to transmit data. [23].

In addition to Sigfox, Semtech also developed the LPWAN technology known as LoRa Technology. The technology is designed for a combination of remote, low power consumption, and secure small-size data transmission. It also operates on an unlicensed SUB-GHz ISM band using a so-called chirp spread spectrum (CSS) modulation to optimize power consumption and broader communications networks. LoRa Technology uses the combination of two layers; the physical layer is known as LoRa for the connectivity and the MAC layer known as LoRaWan.

IV. Network Model

This section describes the proposed network model to study the scalability of LPWAN. The communication model in this study mimics the communication protocol between end-node and gateway for scalability study purposes. The following are assumptions for behaviour of data signal from end-node to be received by the gateway based on [3], [7], [11].

A. The Interference Conditions

In this model, the end-nodes are group into two types which known as reference node and interference node. The reference node is current end-node transmitting data to the gateway at present. While, the interference node refers to others end-nodes beside the reference node that transmitting data signal before, present, or after the reference node transmitting data signal. The received status of data signal for the reference and interruption nodes at gateway can determine whether data signal is successfully received based on the collision condition.

Data signal interruption between the interference and reference nodes are assumed to base on three main parameters; SF, channel and transmission time. If data signals arrive at the gateway from the reference and interference nodes which use the same SF and channel, then all data signal are considered unsuccessfully received by the gateway. The gateway will receive all data signals if the SF and channel used are different. Data signal in this condition is said to be orthogonal to each other. Table I provides detail of the interference condition for both reference and interference nodes.

Data signal interference happens when both of the reference and interference nodes have the same SF and channel. However, data signal can be successfully received by the gateway if both data signal of the reference and interference nodes are being downloaded by the gateway, which passes the preamble time of data signal. Fig. 1 illustrates all possible interference conditions by the end-nodes.

Data signal for the interference node in Case 1 and 6 are successfully received by the gateway as there is no data signal collision with the reference node. In Case 2, data signal from the reference node has successfully received to the gateway. Data signal arrives at the time where the preamble of data signal for the interference node already being downloaded by the gateway. In this situation, both of data signals from the reference and interference node are successfully downloaded by the gateway. This situation is also fit for Case 6, where the roles of the reference and interference nodes exchange. When data signal from the reference node arrives during the preamble of data signal from interference node is being downloaded, both of data signals are assumed to be not received by the gateway as shown in Case 3. This condition is the same as Case 4, and Case 5, where the roles of reference and interference node exchange. Table II shows the received status of the interference and reference node at the gateway.

B. SF Selection

In this study, the SF selections used in the network model for the end-nodes was inspired by [2]. The selection of the SF depends on the distance between the end-node and the gateway. When gateway received data signal from the end-node, it also records the RSSI and SNR value of data signal. Typically, the RSSI and SNR values increase when the distance between the end-node and gateway increases. However, the data signal may attenuate depending on the condition of line-of-sight between the end-node and gateway, which results in increasing the recorded RSSI and SNR values. The end-node requires a higher SF to transmit data signal to the gateway, depending on data signal condition [11].

From previous study, the assumption made for the end-node which located far away from the gateway will use the SF of 12. Data signal is expected to be able to reach the gateway. However, data signal cannot be received by a gateway if the location of the end-node is located too far due to data signal attenuation. It is reasonable to have a limit distance between end-node and gateway for data signal to successfully receive by the gateway. Table III shows the selection of SF value for the data transmission base on the distance between the end-node and gateway. The assumption for the distance between the end-node and gateway is 2 kilometres for each SF. When the distance is over 12 kilometres, the signal is lost and did not received by the gateway.

TABLE. I. INTERFERENCE CONDITIONS OF THE DATA SIGNAL

SF	Channel	Condition
Same value	Same value	Interference
Same value	Different value	No Interference
Different value	Same value	No Interference
Different value	Different value	No Interference

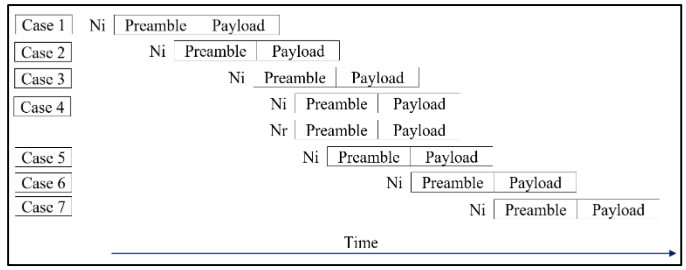


Fig. 1. Diagram of Reference and Interference Node based on Time.

TABLE. II. STATUS OF REFERENCE NODE AND INTERFERENCE NODE

Case	Status Ni	Status Nr
1	Receive	Receive
2	Receive	Receive
3	Loss	Loss
4	Loss	Loss
5	Loss	Loss
6	Receive	Receive
7	Receive	Receive

TABLE. III. SF SELECTION OF THE END-NODE

SF	RSSI (dBm)	Distance (km)
7	(-124, -100)	<2
8	(-129, -124)	2 – 4
9	(-130, -129)	4 – 6
10	(-133, -130)	6 – 8
11	(-135, -133)	8 – 10
12	(-137, -137)	10 – 12
12	(-140, -137)	12 – 14
~	~	>14

C. Gateways Location

In this study, multiple gateways were used to receive the data signal from the end-nodes. The number of gateways used are 2, 4, and 6. The model used 2-dimensional network field with the same length, (L x L). The locations of the gateway were based on the length (L) of the network field. Let say the coordinate of a gateway, $GW(g) = (GWx, GWy)$ where g is the number of the gateway, GWx is the coordinate in x-axis and GWy is the coordinate in y-axis. Then the locations of the gateway were based on the number of getaways used and were given in below equations.

If gateways used are 2

$$GW(1) = (\, L/4, L/2 \,) \tag{1}$$

$$GW(2) = (L - L/4 \,, L/2) \tag{2}$$

If gateways used are 4

$$GW(1) = (L/4 , L/4) \quad (3)$$

$$GW(2) = (L-L/4 , L/4) \quad (4)$$

$$GW(3) = (L/4 , L-L/4) \quad (5)$$

$$GW(4) = (L-L/4 , L-L/4) \quad (6)$$

If gateways used are 6

$$GW(1) = (L/4 , L/4) \quad (7)$$

$$GW(2) = (L/2 , L/4) \quad (8)$$

$$GW(3) = (L-L/4 , L/4) \quad (9)$$

$$GW(4) = (L/4 , L-L/4) \quad (10)$$

$$GW(5) = (L/2 , L-L/4) \quad (11)$$

$$GW(6) = (L-L/4 , L-L/4) \quad (12)$$

V. Simulation

The performance of proposed model is executed via simulation using MATLAB platform. Let say there are *N* numbers of end-nodes distributed randomly in *Lx* x *Ly* two-dimensional network field. The end-nodes are assumed to use specific SF based on the distance between the end-node and the gateway *d*, as discussed in the previous section. Let *D*(*n*) = (*x*(*n*), *y*(*n*)) be the coordinate of the distributed end-nodes and *GW*(*g*) = (*GWx*, *GWy*) be the coordinate of the gateway location. Where *n* = {1, 2, 3…, *N)* and g is the number of gateways. Then, the distance for the end-node *j* from the gateway *g* is defined as;

d(GW(g), D(n)) =

$$[(GWx\ (g) - x(n))^2 + (GWy(g) - y(n))^2]^{1/2} \quad (13)$$

Typically, in LPWAN, one end-node can transmit data signal and is received by multiple gateways. The network will decide which optimal gateway for the next data transmission of the end-node based on the link strength at the gateway [11]. The received signal strength at the gateway is mainly related to the distance between the node and the gateway. The proposed model used these conditions for the end-node to choose the nearest gateway to transmit the data signal. Then, the network will assign the SF based on the distance for the end-node between the end-node and the gateway.

Additionally, the gateway also randomly assigned the channel (CH(n)) for the end-nodes in the range of [1, CH], where CH is the total number of channels. The starting time is assigned randomly for the end-nodes to start transmitting the packet data to mimic the real application of the end-nodes. Starting time (ST(n)) is randomly chosen by the end-nodes based on the range time [0, Application time]. Application time is time for end-node to transmit next data signal after complete transmitting the current data signal.

CH(n) = rand ([1, CH]) (14)

ST(n) = rand ([0, Application time] (15)

Then, the end-node starts to transmit the packet data to the corresponding gateway based on its starting time. When the end-node complete transmitting the packet data, the end-node will set the new starting time (New_ST(n)) with the combination of the starting time, time-on-air (ToA), and processing time (PT). Processing time is time for the end-node to process the data for the next transmission sequel in the range of [0,1.000s]. Note that, the maximum time for end-nodes to process the data is assumed to be 1s. ToA is the time for data signal from the end-node to successfully receive by the gateway. However, it depends on the size of the payload, bandwidth, SF and code rate used by the end-node for data transmission. Refer to [24] for more information on ToA.

PT(n) = [0,1.000s] (16)

New_ST(n) = ST(n) + ToA(n) + PT(n) (17)

At each of the gateway, data signal is successfully received based on the interference conditions as discussed in the previous section. The simulation is run based on round. Each round ends if all the end-node complete transmitting data signal to the gateway. Once the run has reached the designated total run, the simulation stops. Then, the program calculates the percentage of received packet data (PPD). PPD is the percentage of the total received data signal at gateway over the total number of data signal transmission from end-node. Table IV shows the parameter used in the simulation.

The excerpt of the program used in the MATLAB are given below:

```
Randomly place N end-nodes in network field
Randomly set CH, ST and PT
Calculated the distance for each end-node
Set the SF for the end-nodes
Set the end-nodes to the corresponding GW based on the
distance

for round = 1: 50
for Nr = 1: N
for Ni = 1: N
   if CH(Nr) == CH(Ni)&& SF(Nr) == SF(Ni)
      Both data signals not received by GW
      Status received = 0
   else if Case 1, Case 2, Case 6 or Case 7
      Both data signal received by GW
      Status received = 1
      else if Case 3, Case 4 or Case 5
         Both data signals not received by GW
         Status received = 0
         end
      end
   end
Set New_ST for the end-node
end
end
PPD(round) = sum (Status received) / N * 100
end
PPD = sum (PPD)/round
```

TABLE. IV. PARAMETERS FOR THE SIMULATION

Parameters	Values
Size of network field	24000m x 24000m, 48000m x 48000m
Number of end-nodes, N	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000
Number of gateways	2, 4, 6
Packet payload size	50 bytes
Frequency	868 MHz
Bandwidth	125 kHz
Coding rate	4/5
SF	7, 8, 9, 10, 11, 12
Channel	1,8
Application time (s)	600, 3600
Total Round	50

VI. RESULTS AND ANALYSIS

Fig. 2 below shows the example of 500 end-nodes (represented by blank round shapes) with four gateways (represented black round shapes) in the 24000m x 24000m network field. The different colour of the blank round shapes indicates the end-nodes that transmitted data signal to the respective gateway.

Next, Fig. 3 to 5 shows the percentage of the end-node per gateway. The percentage value is calculated based on the average total number of the end-node transmitted to the gateway for 50 rounds. The average number of end-node per gateway varied as the total number of the gateway increased. The end-nodes are located randomly in the network field while the locations of the gateways are fixed. The unbalance numbers of end-node per gateway will affect data signal throughput of the gateway. For example, as shown in Fig. 5, the total number of end-nodes transmitting data signal to gateway 1, 3, 4, and 6 are higher compared to gateway 2 and 5. The higher number of end-node per gateways will have a higher chance for data signal of end-nodes to collide with each other during data transmission.

In the simulation results, the effect of different parameters can be observed on the PPD using the proposed model. Fig. 6 to 9 show the results for PPD using one and 8-Channels with different application time and size of the network field. Overall, the PPD value decreases when the number of end-nodes increases. Observations of this reduction in PPD occur because the amount of data signal from the sensor node rises, resulting in more data signal arriving at the gateways. These increase the chance of data signal to collide with each other's resulting in data signal loss at the gateway.

The overall value of PPD shown in Fig. 6 increases when the number of gateways increases. Increasing the number of gateways will decrease the throughput load by a single gateway. Besides, increasing the number of channel in the network will also increasing the PPD value. Referring to the interference conditions of data signal, data signal with the different channel will avoid the collision.

Meanwhile, the result in Fig. 7 shows a similar pattern as in Fig. 6. Increasing value of application time results in increasing the PPD value. This is shown in Fig. 6 with the application time of 600s, while in Fig. 7 with the application time of 3600s. In a single channel with two gateways, the PPD value of the application time of 3600s gives a similar result to the PPD value when using 8-channel with two gateways with an application time of 300s. The higher application time increases the time difference of the starting time between the end-nodes (refer to equation 15). This increment in time minimizes the number of end-nodes that has the same or similar starting time. Then, the chance for data signal to collide with each other is also reduced.

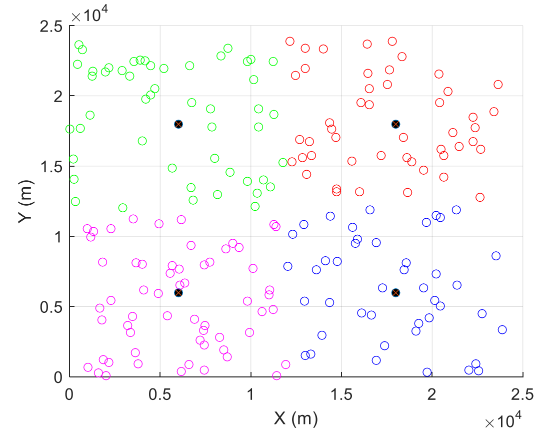


Fig. 2. Location of the End-Nodes and the Gateways.

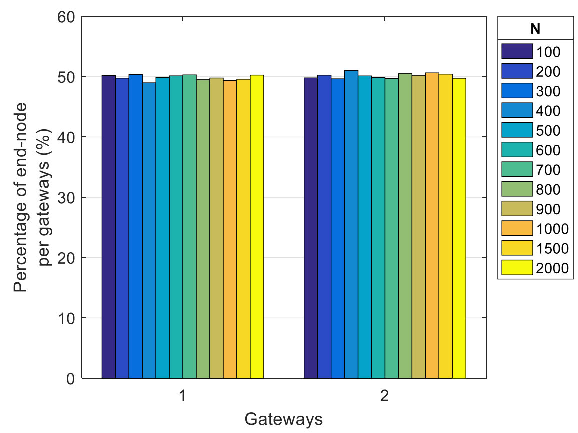


Fig. 3. Percentage of the End-Node Per Gateways using 2 Gateways.

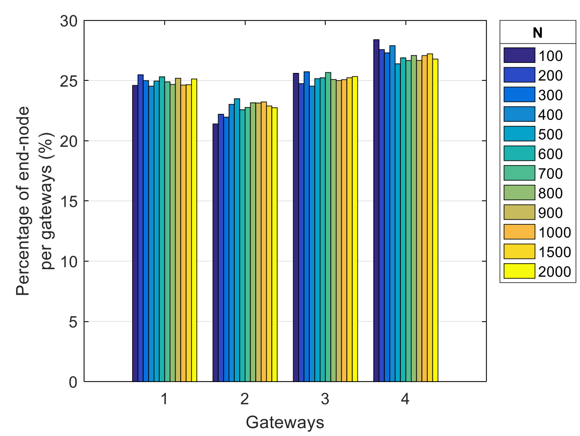


Fig. 4. Percentage of the End-Node Per Gateways using 4 Gateways.

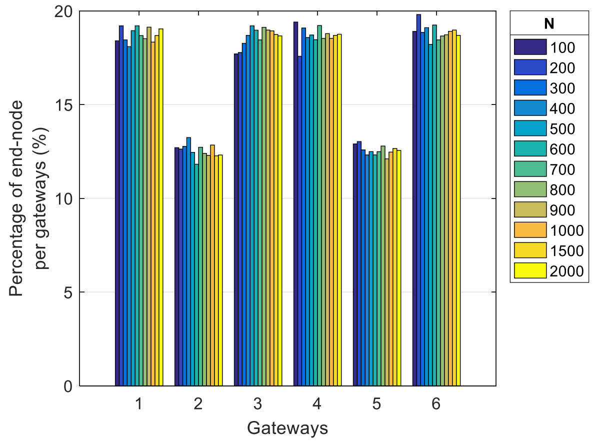


Fig. 5. Percentage of the End-Node Per Gateways using 6 Gateways.

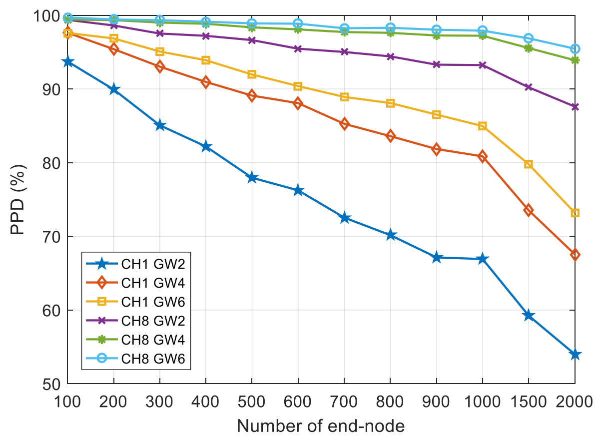


Fig. 6. PPD using Application Time of 600s in 24000mx24000m Network Field.

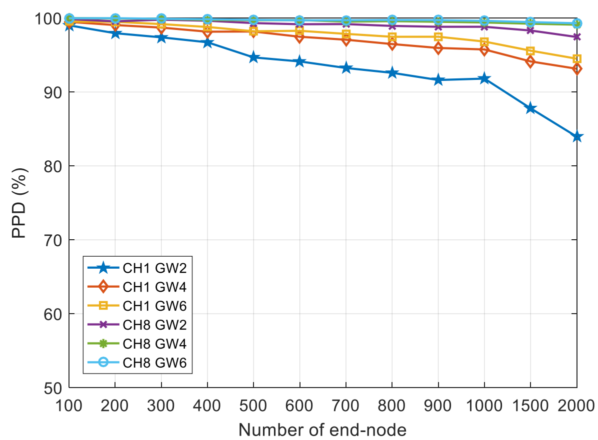


Fig. 7. PPD using Application Time of 3600s in 24000mx24000m Network Field.

Although the PPD value increases when either the number of gateway or channel increases, increasing the channel gives better performance compared to increasing the number of gateways. Meanwhile, increasing the amount of application time gives better results of PPD compared to increasing the number of gateway or channel.

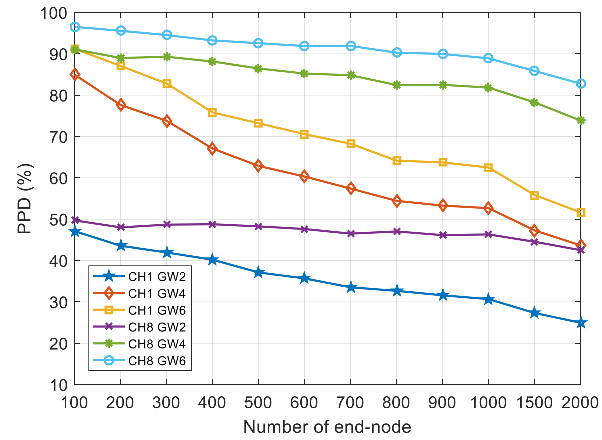


Fig. 8. PPD using Application Time of 600s in 48000mx48000m Network Field.

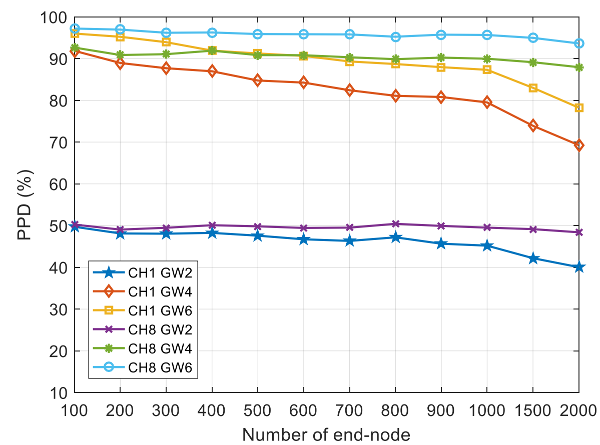


Fig. 9. PPD using Application Time of 3600s in 48000mx48000m Network Field.

Fig. 8, and 9 show the results use the similar parameter as in Fig. 6 and 7 but in double size of the network field. Overall, the PPD value also decreases as the number of end-nodes increases which are similar to the previous results but with a lower PPD value. Increasing the amount of channel or the value application time results in increasing the PPD value. The PPD value when using only two gateways is at 50% and below. This effect is due to increasing the size of the network field. This decrement value of PPD indicates that only half of data signal from the end-node received by the gateways. The gateway solely collected data signal when the location of the one-nodes was in the range of the set limit distance. However, the value of PPD increases when using more gateways to receive data signal.

VII. Conclusion

In this paper, the development and simulation of a comprehensive model of LPWAN to study the scalability using MATLAB simulator is presented. This model includes several assumptions based on the behaviour of LPWAN communication between the end-nodes and gateways, such as the interference conditions of data signal and the selection of the Spreading Factor and the application time. The results show that increasing number of end-nodes, decreases the value of PPD. However, the PPD value increases when the number of gateways, channel and application time increase. The locations of the gateways are directly influencing total number of the end-nodes. The placements of these gateways affect the total number of the end-nodes per gateway. Data signal collisions are more likely to occur when more end-nodes transmitting to a single gateway. Deploying more gateways may overcome this problem. However, in real-time application, increasing the number of gateways will double the cost. Meanwhile, selection of high performance of LPWAN devices is important in order to support high number of channels.

The future scope of the current proposed work can be developed by choosing the optimal locations of the gateways. This optimal location of the gateways should be able to increase the data delivery to the gateways compared to this current proposed location of the gateways when using a similar network environment.

ACKNOWLEDGMENT

The authors would like to be obliged to University Malaysia Terengganu and Ministry of Education Malaysia for providing financial assistance under project Fundamental Research Grant Scheme (FRGS) no. 59466.

REFERENCES

[1] K. C. Serdaroglu and S. Baydere, "WiSEGATE: Wireless Sensor Network Gateway framework for internet of things," Wirel. Networks, vol. 22, no. 5, pp. 1475–1491, Jul. 2016.

[2] I. S. Ismail, N. A. A. Latiff, and N. A. A. Aziemah, "Performance Analysis of Data Recovery via Application Layer for LPWAN," in 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring), 2019, vol. 2019-April, pp. 1–5.

[3] O. Georgiou and U. Raza, "Low Power Wide Area Network Analysis: Can LoRa Scale?," IEEE Wirel. Commun. Lett., vol. 6, no. 2, pp. 162–165, Apr. 2017.

[4] A. Hoeller, R. D. Souza, O. L. A. Ĺopez, H. Alves, M. De Noronha Neto, and G. Brante, "Exploiting Time Diversity of LoRa Networks Through Optimum Message Replication," Proc. Int. Symp. Wirel. Commun. Syst., vol. 2018-Augus, pp. 1–5, 2018.

[5] A. Mahmood, E. Sisinni, L. Guntupalli, R. Rondon, S. A. Hassan, and M. Gidlund, "Scalability Analysis of a LoRa Network Under Imperfect Orthogonality," IEEE Trans. Ind. Informatics, vol. 15, no. 3, pp. 1425–1436, Mar. 2019.

[6] A. Hoeller, R. D. Souza, O. L. Alcaraz Lopez, H. Alves, M. de Noronha Neto, and G. Brante, "Analysis and Performance Optimization of LoRa Networks With Time and Antenna Diversity," IEEE Access, vol. 6, no. c, pp. 32820–32829, 2018.

[7] M. C. Bor, U. Roedig, T. Voigt, and J. M. Alonso, "Do LoRa Low-Power Wide-Area Networks Scale?," Proc. 19th ACM Int. Conf. Model. Anal. Simul. Wirel. Mob. Syst. - MSWiM '16, no. November, pp. 59–67, 2016.

[8] F. Cuomo, M. Campo, A. Caponi, G. Bianchi, G. Rossini, and P. Pisani, "EXPLoRa: Extending the performance of LoRa by suitable spreading factor allocations," Int. Conf. Wirel. Mob. Comput. Netw. Commun., vol. 2017-Octob, 2017.

[9] D. Zorbas, G. Z. Papadopoulos, P. Maille, N. Montavont, and C. Douligeris, "Improving LoRa Network Capacity Using Multiple Spreading Factor Configurations," 2018 25th Int. Conf. Telecommun. ICT 2018, pp. 516–520, 2018.

[10] A. Lavric and V. Popa, "Performance Evaluation of LoRaWAN Communication Scalability in Large-Scale Wireless Sensor Networks," Wirel. Commun. Mob. Comput., vol. 2018, pp. 1–9, Jun. 2018.

[11] J. Haxhibeqiri, F. Van den Abeele, I. Moerman, and J. Hoebeke, "LoRa Scalability: A Simulation Model Based on Interference Measurements," Sensors, vol. 17, no. 6, p. 1193, May 2017.

[12] A. M. Yousuf, E. M. Rochester, B. Ousat, and M. Ghaderi, "Throughput, Coverage and Scalability of LoRa LPWAN for Internet of Things," 2018 IEEE/ACM 26th Int. Symp. Qual. Serv. IWQoS 2018, 2019.

[13] A. Tiurlikova, N. Stepanov, and K. Mikhaylov, "Method of Assigning Spreading Factor to Improve the Scalability of the LoRaWan Wide Area Network," Int. Congr. Ultra Mod. Telecommun. Control Syst. Work., vol. 2018-Novem, pp. 1–4, 2019.

[14] D. H. Kim, J. Y. Lim, and J. D. Kim, "Low-Power, Long-Range, High-Data Transmission Using Wi-Fi and LoRa," in 2016 6th International Conference on IT Convergence and Security (ICITCS), 2016, pp. 1–3.

[15] P. Neumann, J. Montavont, and T. Noel, "Indoor deployment of low-power wide area networks (LPWAN): A LoRaWAN case study," in 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), 2016, pp. 1–8.

[16] D. Kim and M. Jung, "Data Transmission and Network Architecture in Long Range Low Power Sensor Networks for IoT," Wirel. Pers. Commun., vol. 93, no. 1, pp. 119–129, Mar. 2017.

[17] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, "Long-range communications in unlicensed bands: the rising stars in the IoT and smart city scenarios," IEEE Wirel. Commun., vol. 23, no. 5, pp. 60–67, Oct. 2016.

[18] C. Pham, "Low-cost, low-power and long-range image sensor for visual surveillance," in Proceedings of the 2nd Workshop on Experiences in the Design and Implementation of Smart Objects - SmartObjects '16, 2016, pp. 35–40.

[19] J. Petajajarvi, K. Mikhaylov, A. Roivainen, T. Hanninen, and M. Pettissalo, "On the coverage of LPWANs: range evaluation and channel attenuation model for LoRa technology," in 2015 14th International Conference on ITS Telecommunications (ITST), 2015, pp. 55–59.

[20] I. I.S., A. L. N.A., R. F.Z., and A. A. S., A Review on Performances Evaluation of Low Power Wide Area Networks Technology. Springer Singapore, 2019.

[21] A.-I. Pop, U. Raza, P. Kulkarni, and M. Sooriyabandara, "Does Bidirectional Traffic Do More Harm Than Good in LoRaWAN Based LPWA Networks?," no. April, Apr. 2017.

[22] I. LoRa Alliance, "LoRaWANTM 1.0.2 Regional Parameters," San Ramon, CA, 2017.

[23] Sigfox, "Radio Technology Keypoints." [Online]. Available: https://www.sigfox.com/en/sigfox-iot-radio-technology. [Accessed: 09-Jan-2018].

[24] Semtech, "SX1272/3/6/7/8: LoRa Modem Design Guide," 2013. [Online]. Available: https://www.semtech.com/uploads/documents/LoraDesignGuide_STD.pdf. [Accessed: 09-Apr-2018].

Tuberculosis Prevention Model in Developing Countries based on Geospatial, Cloud and Web Technologies

Innocent Mwila[1]
Department of Electronics and Electrical Engineering
University of Zambia
Lusaka, Zambia

Jackson Phiri[2]
Department of Computer Science
University of Zambia
Lusaka, Zambia

***Abstract*—Information is important when making decisions. Decisions which are based on gut feeling and made in the absence of evidence always tend to be less effective in most situations. This is also the case when it comes to Tuberculosis (TB) disease control and prevention intervention planning and implementation. The lack of evidence-based information upon which decisions for action to help with the prevention of spread of TB has proved to be less effective in the prevention of the disease as TB keeps spreading. The aim of this paper was to design and develop a prototype system that would provide TB program managers with information and tools which can be used to make decisions which can effectively influence the fight against the spread of TB through the application of cloud computing, geospatial data analysis and web technologies. The system would improve disease monitoring and tracking through the use of the identified technologies, by displaying the geographical distribution of TB cases in the communities on a mapping application as well as providing reports which TB program managers can use to make decisions when planning and implementing disease control and prevention activities.**

***Keywords*—Evidence-based; monitoring; cloud computing; geospatial data analysis; mapping; web technologies; information; decision-making; tuberculosis prevention**

I. Introduction

TB is one of the leading infectious diseases in the world with an annual average of 10 million cases being reported between 2016 and 2018, and causing an average of 1.4 million deaths in the same period which translates into 25% of deaths globally [1][2][3][4] and therefore, making TB the number one cause of death by a single infectious agent [1]. TB caused 17,000 deaths in Zambia in 2015 [5] and an average of 36,000 cases have been recorded annually between 2016 and 2018 in the country [6]. Zambia is among the 30 high TB burden countries in the world [7]. TB is an infectious disease which spreads from an infected person to the non-infected, and one way to prevent the spreading of the disease is by identifying individuals who are infected and then put measures in place to curtail the spread, to protect the ones not infected and Information and Communication Technologies (ICTs) can be utilized for this purpose, to improve surveillance [8].

ICTs play a significant and cornerstone role in the development of countries and the world at large [9] and the ability to make decisions, do business and communicate in this information age dictates the rate at which social, political and economic development takes place as well as the advancement towards epidemic control [10]. ICTs have been used in different ways to improve disease surveillance [11][12] and thereby preventing further spreading or at least reducing the rate at which some diseases are spread.

ICTs have been identified as an important enabler in the fight against diseases and TB is not an exception [13][14]. Disease surveillance is the continuous, systematic monitoring, tracking, collection and analysis of data and the provision of information which leads to decisions being made and action taken to prevent and control infectious diseases [15]. The use of modern technologies has improved disease surveillance.

ICTs, particularly cloud computing, geospatial data analysis and web technologies have been less utilized in TB monitoring and tracking in Zambia. This has resulted in little being known about the geospatial distribution of TB cases and untimely to non-availability of other information about the disease. Information is not readily available when required for decision making and as such it is unclear as to which areas and communities TB program managers should urgently direct the resources in the fight against TB. This paper looks at how cloud computing, geospatial data analysis and web technologies can be applied to empower TB program managers with information needed to make decisions as they plan and implement disease control and prevention activities.

The aim of this paper was to design and develop a prototype system that would provide information which TB program managers can use for decision-making to help in the prevention of the spread of TB, through improved monitoring and tracking using cloud computing, geospatial data analysis and web technologies.

Firstly, the paper established factors which escalate the spread of TB in Zambia and then designed a model that would be used to monitor and track the spreading of TB in the communities based on cloud computing, geospatial data analysis and web technologies, and finally, based on the designed model, developed a prototype system to help with decision-making in the fight against the spreading of TB.

The remaining part of the paper is divided into six sections as follows; The literature review is given in Section II while the related works are covered in Section III. In Section IV, the paper gives the details on the research methodology. The results are given in Section V, the discussion in Section VI and finally the conclusions and recommendations are given in Section VII.

II. Literature Review and Related Works

In disease surveillance, the speed at which and the accuracy with which diseases or outbreaks are reported draws a difference between a successful or a failing disease control and prevention campaign. Leveraging the ability of ICTs such as real time reporting with the ability of Geographic Information Systems (GIS) to identify and display the geographical position of objects are being used to improve disease surveillance and this has resulted in improvements in disease prevention activities. The review looked at the TB disease, how it is caused, the factors which promote the spread of the disease, and how cloud computing, geospatial data analysis and web technologies have been used to support decision-making in countries such as India and South Africa to prevent the of spread TB through improved monitoring and tracking. The review also looked at some related works which have been done in the recent past in the health sector.

A. Tuberculosis and its Transmission Factors

TB is an airborne disease which mainly affects the lungs and can spread to other parts of the body [1]. It is caused by the Mycobacterium Tuberculosis bacterium and spreads from one individual to the other by the exhaling of the bacteria from the infected person's lungs and the subsequent inhaling of these bacteria by a non-infected individual [1][16]. The disease is preventable and being aware of the factors which lead to its spread is key in stopping further spread. Fig. 1 shows some of the TB transmission risk factors.

Susceptibility is the likelihood of one to catch a disease due to a weakened immune system [16] and the weaker the immune system the higher the likelihood of one to get infected with TB when exposed to the TB causing bacteria. The amount of the TB bacteria which an infected person can expel into the air when they cough or otherwise is known as infectiousness [17]. Environment factors such as the concentration of infectious droplets in the air, the exposure to the TB infected individuals in small enclosed spaces and inadequate ventilation [16][17] are among the prominent drivers in the spread of the disease. Ventilation is the flow of air in and out of a space to get rid of stale particles from that space [18]. Inadequate ventilation results in insufficient dilution or removal of infectious TB bacteria from the environment [16][17] and this supports the transmission of TB. Proximity and length of exposure to the TB causing bacteria are other factors pointed out by [16][17] to be key in the spread and transmission of TB, as longer duration of exposure to a person with infectious TB accelerates the risk of transmission of the disease to the people around that person. Proximity and prolonged exposure to the TB causing bacteria in homes and other settings is mainly promoted by crowding [16][17]. A house is said to be crowded if the ratio of the number of its occupants to that of the total number of rooms of that house is greater than or equal to 1.5 [19][20].

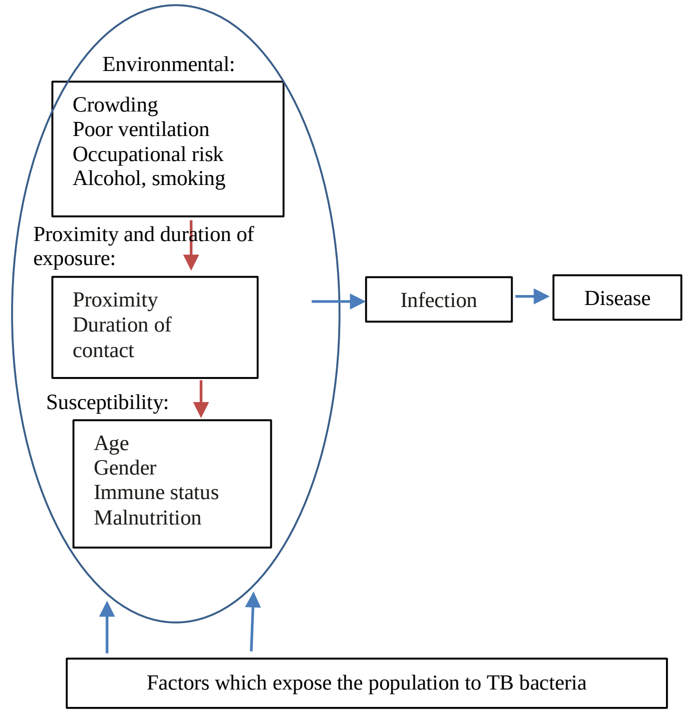


Fig. 1. TB Transmission Risk Factors [16][17].

Influencing the disease transmission factors which we have control over is important in the fight against the spread of TB and much of the emphasis in this fight must be placed on altering or being aware to these factors.

B. ICTs in Disease Survaillence and Spread Prevention

ICT is the capturing, processing, storing and transmission of information electronically using a variety of communication methods and devices such as satellite networks, mobile phones, wired and wireless networks as well as other digital forms [10]. The prevention of the spread of TB can be enhanced through better disease surveillance and ICTs can be harnessed for this purpose. ICTs have been used in different ways to improve disease surveillance [21][12] and thereby preventing further spreading or at least reducing the rate at which some diseases are spread. Using ICTs, in particular, the leveraging of cloud computing, improved data management (through database technologies), geospatial data management and web technologies in TB disease tracking and monitoring can help to prevent the spread of new cases or reduce the rate at which the disease is spread as has been the case in India, United States and Ireland [22][15].

C. Geo Spatial Data Analysis

Spatial data indicates where objects are within a given coordinate system while geo is the absolute or relative position of an object on the Earth's surface. Geospatial data, therefore, represents the positioning of an object on earth's surface and is associated with a coordinate system and so can be transformed onto a map. Vector data is used to depict a given point on a map and is in form of latitude and longitude pairs, which show the location of objects on a map and can be analyzed [23] and used to action decisions. The scrutinizing of spatial data to interpret the relationship of objects and other associated attributes on the earth's surface is known as Geospatial Data Analysis [23] and Geographic Information Systems (GIS) are software systems used to capture, store, retrieve, analyze and display spatial data [24]. The advantages of GIS which includes cost effectiveness, handling of large volumes of spatial data and the ability for real time reporting have made GIS to be widely applied in disease surveillance and decision-making [25].

D. Cloud Computing

Cloud computing is the use of remotely hosted servers for data management instead of local computers and servers [26]. The servers, other computer resources, hardware and software are hosted offsite and made available via the internet by service providers. Cloud computing offers several advantages such as the pulling of resources to utilize economies of scale and scalability, cost efficiency, device independency, high computing power, location independency, redundancy and improved availability [26] while allowing users to use low specification computers to take advantage of maximum computing capabilities.

Cloud computing has presented itself as tool that can be harnessed for disease surveillance and help with the prevention of the spread of diseases by providing early warning systems for infectious diseases so that timely steps can be taken to avoid epidemic breakout [27]. Whereas traditional systems of disease monitoring are time consuming, slow and use less efficient data collection and monitoring systems, cloud computing provides speed, accuracy and efficiency in the epidemic outbreak detection using intelligent systems, databases, improved analytics, modeling, visualization and ontology mapping offering real time options for decision-making [27].

E. Disease Survaillence using Web Technologies

Web technologies collectively describe the types of technology used in communication over the Internet. The web of connections enables people to communicate and access files and documents located on other computers (and other electronic devices) in the network using web browsers. A web browser is a program that runs on client computers and requests for information, files, documents or services from other computers [28]. Examples of web browsers include Internet Explorer, Firefox, Chrome and Opera. The ability of web technologies to report disease occurrences in real time, at low cost and the simultaneous access to the data and other reports by various team players in disease surveillance regardless of one's location as long as internet connectivity is available are motivating factors. Although there are challenges which web-based applications are associated with, such as inaccurate information interpretation and privacy issues, it is recommended that web applications are applied in disease surveillance as the strengths, opportunities and advantages by far outweighs the challenges of tradition surveillance systems which are often expensive and inefficient due to inherent delays between disease occurrence and notification [29][30]. Quick and strategic decision-making are important in disease monitoring and enhanced surveillance can be made possible using web technologies [30] in epidemiological investigations as well as management and protection of exposed individuals.

F. Mobile Applications in Disease Survaillence

The convenience of being able to capture data from any place and send it to a database in a remote location using a device that one can move with to any place of choice makes mobile applications very handy for data collection [31][32] especially in disease monitoring and tracking. A mobile application is a software system that is programmed for and runs on a small movable handheld device and is accessible from any place.

Mobile applications are programmed for operating systems such as Android, Symbian and Windows, and they run on smart phones [33]. Main advantages of mobile applications include quick and real time communication, and less computation which results in less power consumption [31][34], which makes them a reliable option when power is a constraint as is the case in many developing countries. Mobile applications have been used in disease surveillance and have demonstrated reliability, accuracy and creativity upon which players in the health sector can rely for infectious disease monitoring and tracking [32][35].

III. Related Works

A. South Africa

Spatial data analysis was used in a research conducted by [36] in South Africa to determine the geographical distribution of TB cases in Western Cape's suburbs with the highest incidences of the disease. GIS was used to determine the distribution of the TB cases and the results of the research study showed that the TB cases were unevenly distributed in the communities with 1,835 out of the 5,345 (34.3%) residential homes housing at least 1 case of TB in the past decade while 3 or more cases occurred in 483 houses.

The study showed that the TB cases in a high incidence community spread unevenly though cases repeatedly occur in some households. The study further showed that there were higher incidences in households with the smallest plots in the communities, suggesting overcrowding. The lessons learned from this study were used to concentrate TB prevention services in specific areas as informed by GIS technologies.

B. India

The aim of the study under review was to identify geographical clusters with the prominent number of TB cases to design interventions of dealing with the escalating incidence of TB. In this study, reference [37] used spatial scan statistics to identify purely spatial and space-time clusters of TB.

The research was conducted in Almora District, Uttaranchal State, India. A spatial scan statistics software program was used to test the presence of statistically significant spatial as well as space-time clusters of TB and to identify their approximate locations. The geospatial analysis was done via GIS technologies with input data used being digital maps provided by the local government agency. The results of the study identified TB hotspots in the district. GIS eases the analysis and clearly displays the spatial patterns of disease distribution and can be used to provide insight into the extent and distribution of TB and more importantly, let TB program managers use this information in the allocation of resources to fight the disease, intensify remedial measures and plan out future strategies for impactful TB control.

C. China

Timely information provision was an important factor in this study whose objective was to use mobile applications and Web GIS for decision support to respond to infectious disease emergences. Reference [33] used geo data to display infectious disease emergencies via a Google Mapping API and the information collected by data capturers in the field using mobile phone applications was being shared via a 3G wireless network with decision makers in their respective workplaces who would then make decisions about the course of action in the fight against the spread of the infectious diseases. In this study, baseline data, geo coordinates of locations where infectious disease occurred and other disease parameters were captured throughout China, sent to the server and then the Web GIS application on the computer of the decision makers would process and summarize the data into reports to be used to effect action.

The architecture of the system under review, shown in Fig. 2 had two ends, the data collector interface sitting on an application of a phone or mobile computer and the decision maker interface, on a web browser of a computer.

The reviewed work illustrates how mobile and web-based applications can be used in disease surveillance to enhance decision-making in the face of infectious diseases and [33] recommends the use of the technologies used in this study for disease investigations especially in developing countries.

D. eIDSR

In the run up to and during the early stages of the 2014 to 2016 Ebola outbreak in Sierra Leone, the national health surveillance system, Integrated Disease Surveillance and Response (IDSR) was based on texting, calling or hand delivery of disease surveillance reports from health facilities to the district office for entry into the national system [38]. However, as the Ebola outbreak was peaking, this business model became inadequate as Ebola cases and other diseases' information needed to be availed in real time and accurately in order to allow prevention officers to act and prioritize their activities based on the information, and this led to the designing and development of electronic Integrated Disease Surveillance and Response (eIDSR) system.

eIDSR uses mobile devices for data entry from communities and health facilities, and the data is sent to the web-based national database which runs on DHIS2.

This system helped Sierra Leone to respond to the Ebola outbreak efficiently with reporting rates improving, from less than 40% before eIDSR deployment to greater than 97% with the error rate dropping by 45%.

E. eHealth Cloud

eHealth Cloud is a cloud-based health care system that was developed in Bangladesh to heterogeneously connect the patients, doctors, the government and other players in the health care system to replace the paper-based system [39]. The lack of electronic medical information repository, the inability for doctors to see patient medical history, the impracticality of analyzing medical data and the absence of a heterogeneous communication platform among the health system players were major issues with the paper-based system.

A three tier model as shown in Fig. 3 on page 5 which was designed to address the problems noted in the paper based system has a client and server, and in between them is logic layer which implements the application logic rules and other system functions such as data querying and transfer as well as handling data security issues for eHealth Cloud.

eHealth Cloud offers the unified connected solution for all health care players and allows for seamless interaction and information sharing in the health care system with the capability of storing large volumes of medical data such as laboratory results, diagnostic information and prescriptions.

The front end of eHealth Cloud was developed using the Rich Internet Application platform and is accessible from both desktop and mobile devices while the cloud (database) server was implemented using SimpleDB. Representational State Transfer (REST) was used to link the front end to the cloud server which does all the processing so as to take the processing burden off the clients of eHealth Cloud.

F. FluMob

With the increasing cases of Influenza in Singapore, there was need for a real time monitoring system for tracking and reporting the disease and Flumob system was developed as a response to this need [40]. Flumob is used by health care workers to report influenza cases in real time using a mobile interface and web browsers of computers to the central server. Flumob generates reports which are analyzed and used for disease control and prevention purposes.

Flumob's nonfunctional requirements implementation are Windows server 2008 R2, Apache, PHP, MySQL, Android studio, and xCode for iOS development in terms of software, and a central server machine with four cores Core2 Intel Xeon Processor, 8 GB of RAM, and 500 GB of storage space.

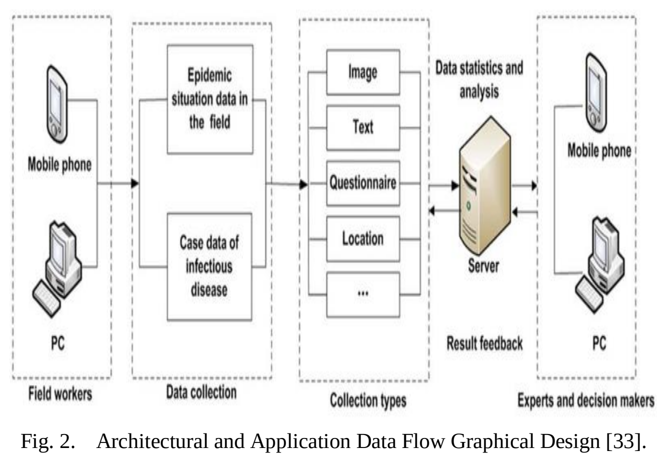


Fig. 2. Architectural and Application Data Flow Graphical Design [33].

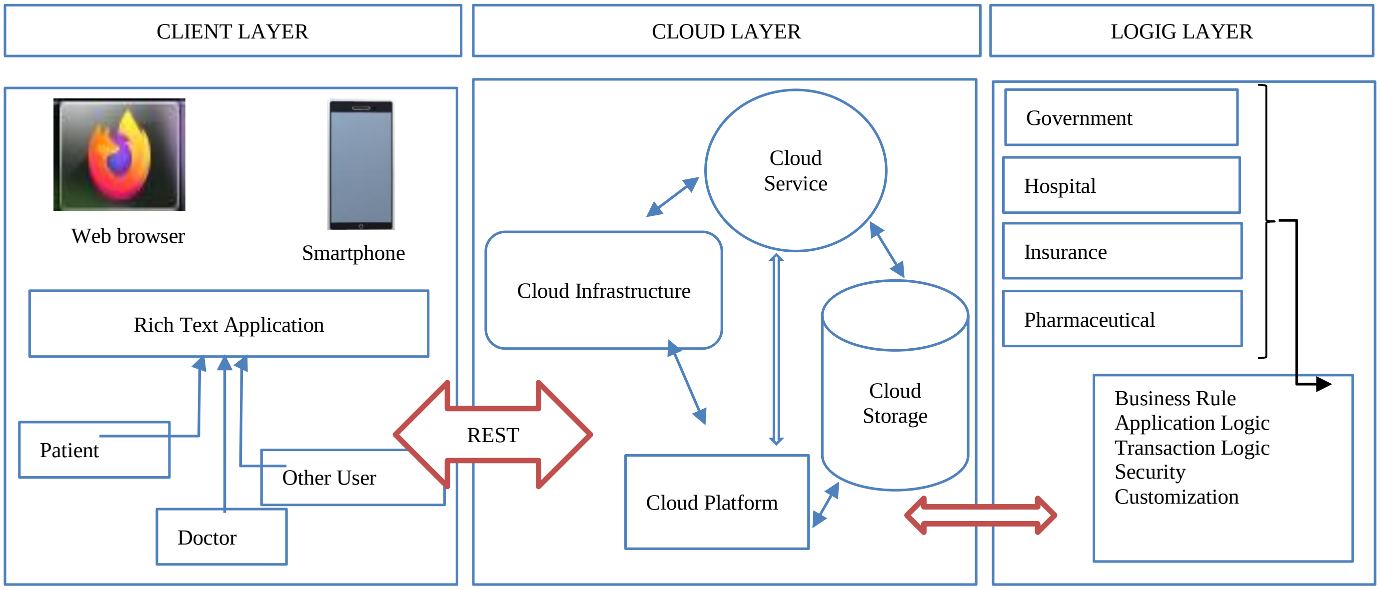


Fig. 3. Three Tier cloud System Design [39].

IV. Methodology

The aim of this paper was to design and develop a prototype system that would provide information which TB program managers can use for decision-making to prevent the spread of TB.

Firstly, the paper established factors which escalate the spread of TB in Zambia using a baseline study and then designed a model that would be used to monitor and track the spreading of TB in the communities based on cloud computing, geospatial data analysis and web technologies, and finally, based on the designed model, developed a prototype system to help with decision-making in the fight against the spread of TB using the Cloud Computing Architecture, Google Mapping API for displaying reported TB cases on a map, PHP for developing the web application frontend and MySQL as the data management platform.

A. Data Collection and Population Study Sample

The baseline study was carried out in Lusaka district of Lusaka province of Zambia. TB patients on TB treatment formed the sample study population. Participants of the baseline study were identified from clinical records at health facilities from which they were obtaining their TB drugs and randomly picked. The health facilities from which participants of the baseline study were recruited included Chawama Level One Hospital, Chipata Level One Hospital, Kalingalinga Clinic, Kanyama Level One Hospital and Mtendere Clinic Table I. shows the five health centers of Lusaka district with the respective number of participants.

The questionnaire for interviewing participants for the baseline study data collection was built around the transmission risk factors identified in the literature review (under Tuberculosis and Its Transmission Factors) and data collection was done using KoboCollect application running on Android as this allowed for the capturing of geo coordinates of participants' homes.

TABLE. I. Population Study Sample

Facility Name	Participants
Chawama First Level Hospital	31
Chipata First Level Hospital	59
Kalingalinga Clinic	25
Kanyama First Level Hospital	81
Mtendere Clinic	49

B. System Modelling

a) Current Business Model: Currently, the TB prevention routine is reactive and is mainly activated once a confirmed bacteriological TB case has been identified. The prevention routine in this model includes a contact tracing activity in which members of the household in which the bacteriologically confirmed TB case has been identified being screened for TB symptoms. In cases where funds are available, TB prevention activities are conducted in these communities in a one fits all manner, for example distribution of literature which teaches coughing etiquettes – the best way of coughing so as not to spread TB. During specific times of the year such as the TB month, which is commemorated in March annually, mass education about the disease is done, for example, through sending of text messages to people on a certain mobile communication network. These messages are generic and are not informed by any evidence. The flow chart in Fig. 4 shows the summary of the current TB prevention model in the communities.

b) Proposed Business Model: The proposed business model took the steps in the current model and incorporated the use of ICTs such as cloud computing, geospatial data analysis and web technologies to capture, store and process data. TB is preventable and being aware of the factors which lead to its spread is key in stopping the further spread. Influencing factors such as proximity and exposure as well as other environmental factors which we have control over is important in the fight against the spread of TB and much of the emphasis in the fight must be focused on altering or being aware to these factors and spatial data analysis of already existing TB cases using GIS can help identify the geographical patterns of transmission and spread of the disease and therefore inform strategic decisions in the fight against TB. Once the data has been processed into maps and reports, TB managers can then use this information to plan and implement TB prevention activities in the communities.

The proposed model seeks to establish the likely causes of the given TB case, this process also involves the assessment of the patient's homes. This information is important when designing prevention activities. The likely causes of the disease are captured together with the details of the patient's house as well as the geo coordinates of that house and sent to the database in the cloud. The web application running on the TB program managers' computer retrieves maps showing the geo distribution of TB cases in the communities and other reports upon which disease prevention and control activities can be planned and implemented. The flow chart in Fig. 5 on page 7 shows the proposed business model.

C. *Application Design*

a) System Architectural Design: From the proposed business model, the architectural design of the system is as shown in Fig. 6 and works as illustrated; the system consists of the data capturing application which sits on the mobile device (Smartphone, Table, Laptop etc.) and is accessible via a web browser. The user captures data (including geo coordinates of patients' homes) on the mobile device and sends the data to the database in the cloud for storage. The web application on the decision maker's computer, embedded with a GIS application is used to display geo locations of TB cases and to generate summary reports which can be used for strategic decision-making by TB program managers to help with TB prevention.

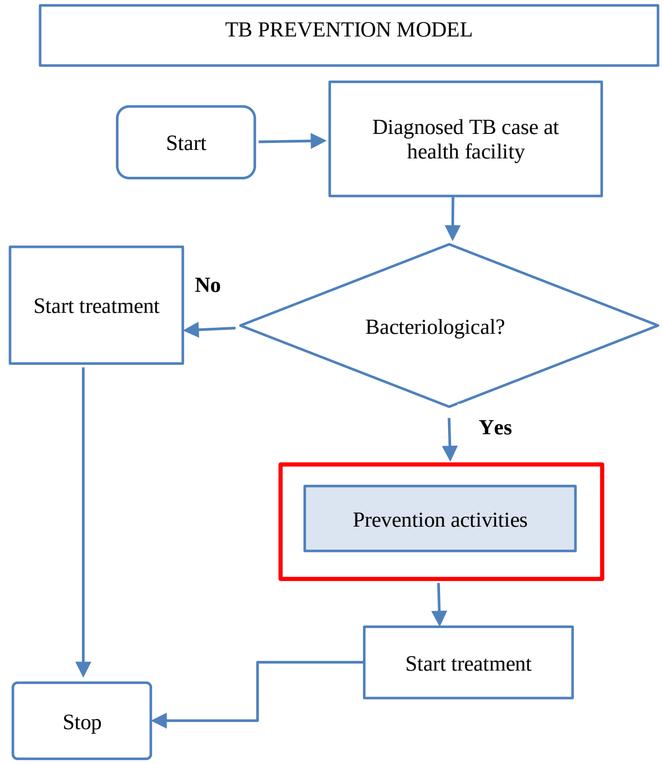


Fig. 4. Current Business Model Flow Chart.

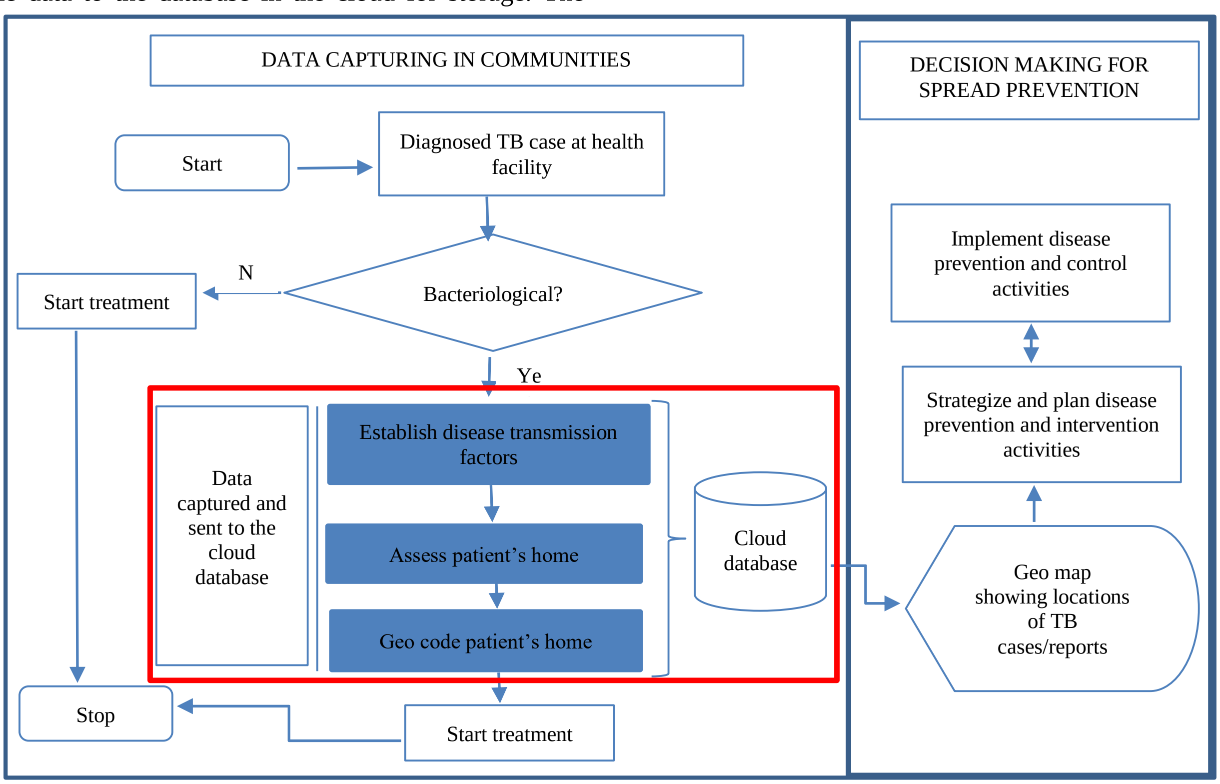


Fig. 5. Proposed Business Model Flow Chart.

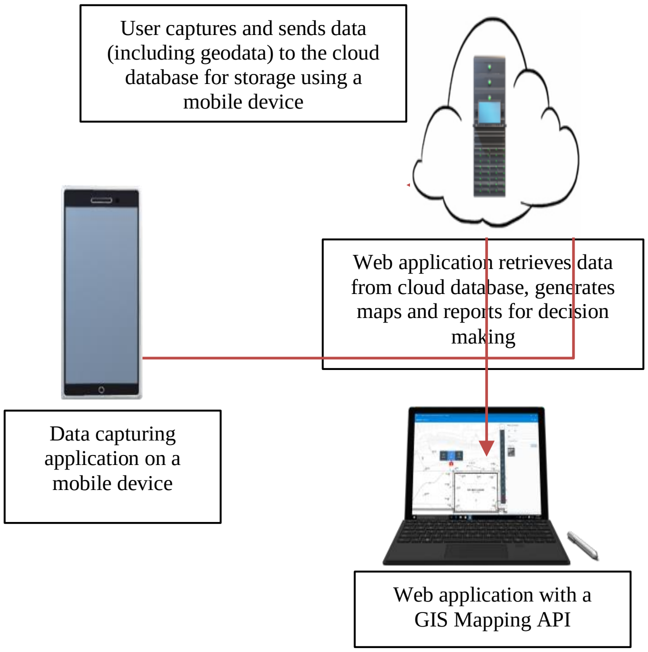


Fig. 6. Architectural System Design.

b) Use Case Diagrams: The system has two actors, the data capturing personnel, to capture data and the TB program manager, who is the system administrator. The TB program manager views maps, runs summary reports and makes decisions about disease control and prevention activities based on the availed information. The TB program manager being the system administrator can also create user accounts for other system users. Fig. 7 and Fig. 8 respectively shows the use case diagrams for the data capturer and TB program manager.

c) Sequence Diagram: Using the architectural system design (Fig. 6) and the two use case diagrams (Fig. 7 and Fig. 8), the logical sequence of activities of the system prototype can be summarized as shown in the sequence diagram in Fig. 9 on the next page.

D. Data Design

MySQL is the Relational Database Management System (RDBMS) that was used to develop the data management platform for the prototype. MySQL is used to store, process and retrieve the data in the system. The data design of the application consists of seven tables, each table being the entity with its attributes. Fig. 10 shows the conceptual design of the system prototype database in the form of an Entity Relationship Diagram (ERD). It shows the database entities.

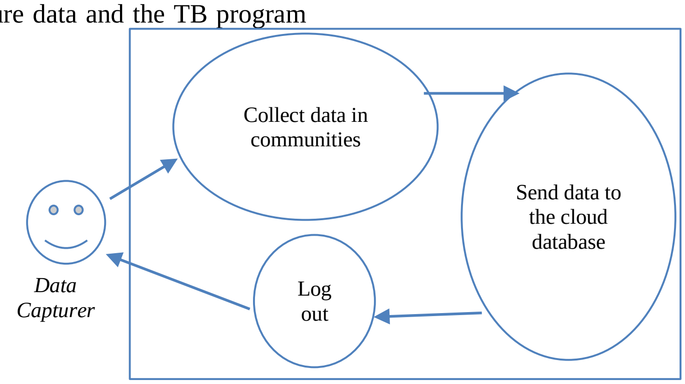


Fig. 7. Data Capturer use case Diagram.

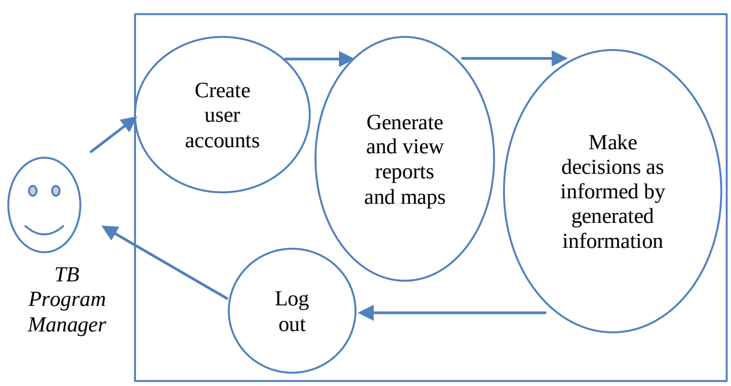


Fig. 8. TB Program Manager use Case Diagram.

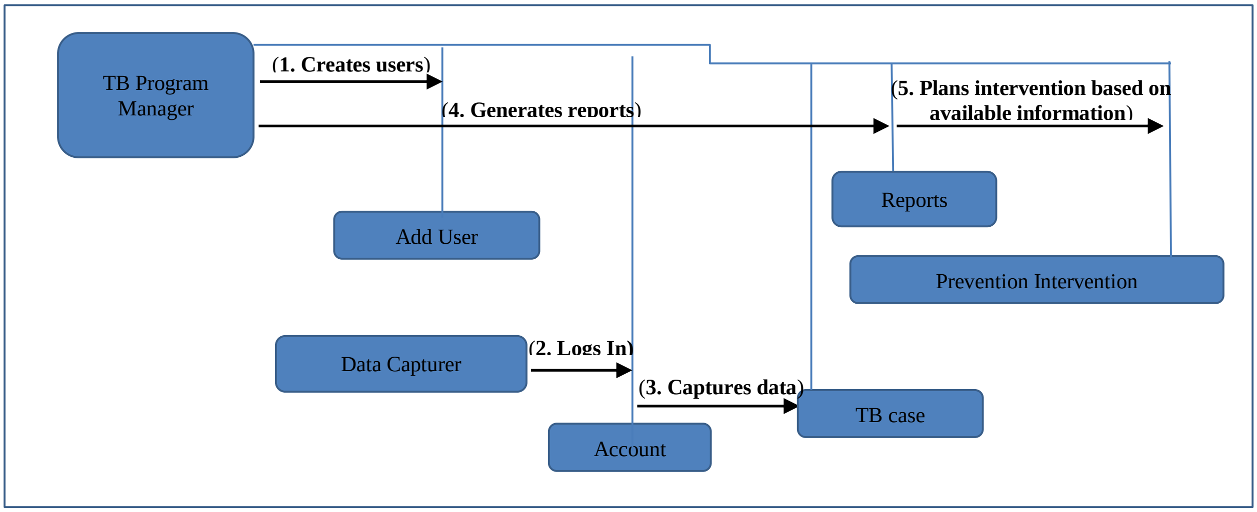


Fig. 9. System Sequence Diagram.

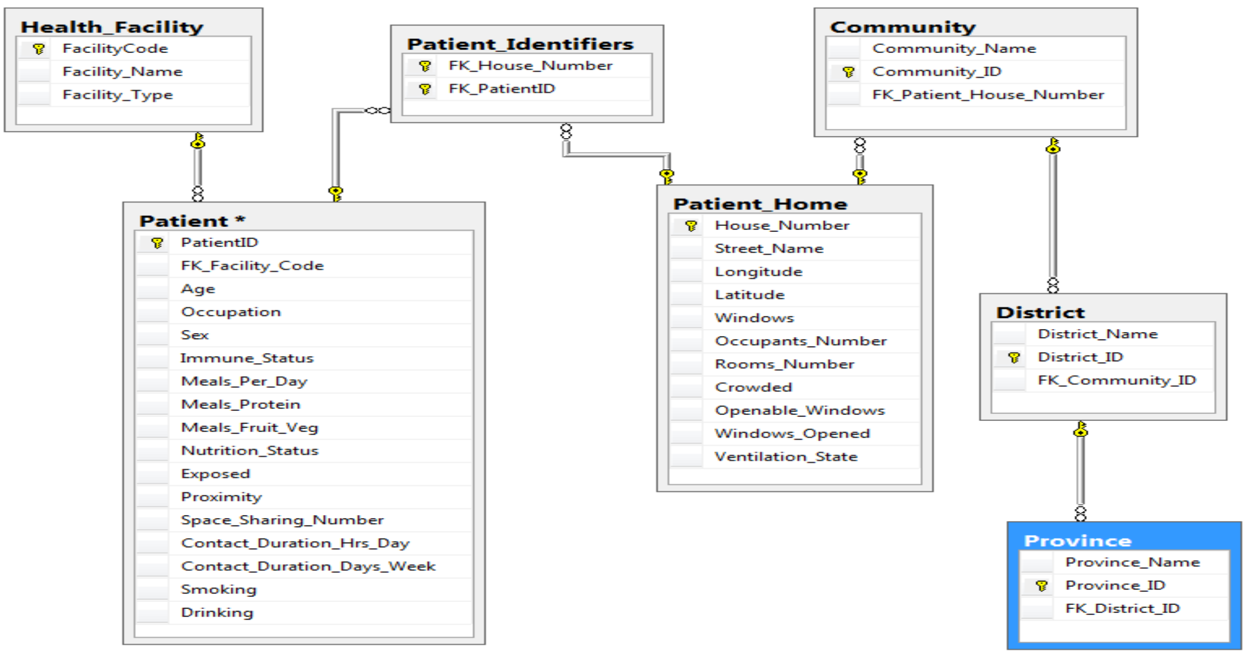


Fig. 10. ERD of the System Prototype.

The relationships depicted in the ERD can be summarized as;

- A Health Facility offers diagnostic and treatment services to Patients.
- Patients live in a Home.
- A Home is located in particular Community.
- A Community is within a District.
- A District is in a given Province.

V. Results

The first section looks at the results as the output of the research activity. It starts by highlighting the results of the baseline study using descriptive statistics and regression analysis. The analysis was done using SPSS and results have been disaggregated by some of selected TB transmission factors identified in the literature review and how likely they are to escalate the spread of TB in Zambia, Lusaka district in particular. The second section shows some modules of the developed prototype system and also highlights the requirements needed for the prototype system to function.

A. Baseline Study Results

The baseline study had 245 participants drawn from five TB diagnostic and treatment health centers around Lusaka district. Table II shows the Frequency Table of distribution of participants across the five health facilities. The results of the survey show that 82% of participants were aged between 20 and 50 years old as shown in Table III while 65% of participants were male as shown in Table IV which displays the disaggregation of participants by sex.

TABLE. II. PARTICIPANTS' FREQUENCY TABLE

Number	Participants distribution			
	Facility	***Frequency***	***Percent***	***Cumulative Percent***
1	Chawama Level One Hospital	31	12.7	12.7
2	Chipata Level One Hospital	59	24.1	36.7
3	Kalingalinga Clinic	25	10.2	46.9
4	Kanyama Level One Hospital	81	33.1	80
5	Mtendere Clinic	49	20	100
	Total	**245**	**100**	

TABLE. III. PARTICIPANTS BY AGE GROUP

Group	Participants age groups			
	Age group	***Frequency***	***Percent***	***Cumulative Percent***
1	<20	9.4	9.4	9.4
2	20-29	85	34.7	44.1
3	30-39	77	31.4	75.5
4	40-49	40	16.3	91.8
5	50-59	13	5.3	97.1
6	60-69	4	1.6	98.8
70+	3	1.2	100	100
	Total	**245**	**100**	

TABLE. IV. PARTICIPANTS BY SEX

Sex ID	Participants by sex			
	Sex	***Frequency***	***Percent***	***Cumulative Percent***
1	F	85	34.7	34.7
2	M	160	65.3	100
	Total	**245**	**100**	

a) Exposure Due to Proximity: The only way TB can be transmitted from the infected person to the one not infected is by exposure to the TB causing bacteria and this exposure is mainly due to being with someone who has active TB. Out of the 245 participants surveyed, 134 remembered being and sharing space with someone who was either on TB treatment or presented with the TB symptoms such as prolonged coughing, fever, chest pains or was unintentionally losing weight during the last one year prior to being diagnosed with TB. The bar graph in Fig. 11 shows that 55% of the respondents could have contacted TB due to exposure to the TB causing bacteria as a result of being in the same space and spending time with someone who had TB.

Linear regression analysis results in Table V show the P Value of $p < .001$, implying a significant relationship between exposure to the TB causing bacteria and TB cases. The coefficient of 23 and 111 being the constant of the linear regression equation means that TB cases will occur if individuals are exposed to the TB causing bacteria.

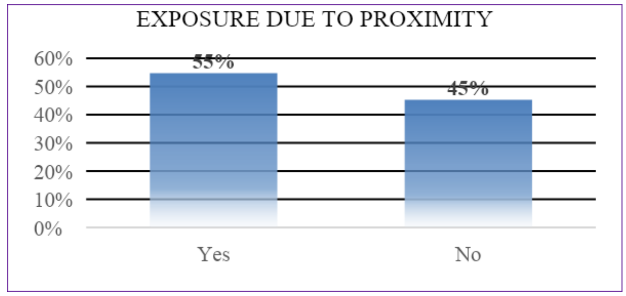


Fig. 11. Exposure Due to Proximity Graph.

b) Ventilation: While a number of participants were exposed to the TB bacteria due to proximity, the problem of TB spreading could have been compounded by poor ventilation with 165 (which is 67%) of the 245 participants' homes which were surveyed having poor ventilation.

Linear regression analysis results show a perfect negative coefficient correlation, -1.000 between ventilation and TB cases. The linear regression equation has -85 as coefficient and 165 being the constant, implying that TB cases are likely to occur as long as ventilation is poor. Table VI shows the regression analysis results of the baseline survey on ventilation.

c) Crowding: The results of the baseline study showed that crowding was eminent with 202 of the surveyed participants' houses falling into the crowded category, translating into 82%. Only 43 out of the 245 houses were not crowded. With exposure to the TB causing bacteria and poor ventilation being common, crowding is likely to escalate the spread of TB.

The Linear regression analysis results show the significance of .000 with coefficient of 159 and 43 being the constant of the linear regression equation and 1.000 as the correlation coefficient, implying a perfect positive correlation between crowding and the spreading of TB. The bar graph in Fig. 12 and the linear regression analysis results in Table VII summarizes the baseline survey results on crowding.

TABLE. V. EXPOSURE VS TB CASES REGRESSION ANALYSIS COEFFICIENTS OUTPUT

Model	**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
	B	***Std. Error***	***Beta***		
(Constant)	111	0		1931292773.353	.000
Exposure	23	0	1	-233920361.389	.000

a. Dependent Variable: TB Cases

TABLE. VI. VENTILATION VS TB CASES REGRESSION ANALYSIS COEFFICIENTS OUTPUT

Model	**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
	B	***Std. Error***	***Beta***		
(Constant)	165	.000		1170045913	.000
Ventilation	-85	.000	-1.000	-427078195	.000

a. Dependent Variable: TB Cases

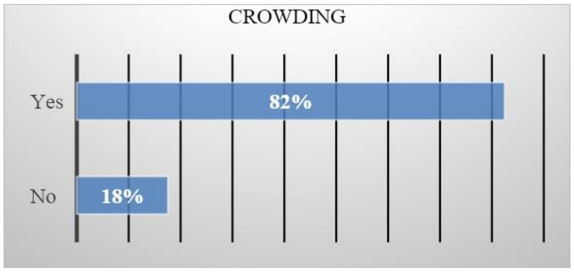


Fig. 12. Results of the Crowding Survey.

TABLE. VII. CROWDING VS TB CASES REGRESSION ANALYSIS COEFFICIENTS OUTPUT

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	***Std. Error***	***Beta***		
(Constant)	43	.000		231471142.26	.000
Crowding	159	.000	1.000	603979776.00	.000

a. Dependent Variable: TB Cases

d) Nutrition: Good nutrition is a deterrent to the development of active TB. When exposed to the TB causing bacteria, a person with a bad nutrition status is highly likely to develop active TB compared to one who has a good nutrition status. The results showed that 92 (38%) participant's nutrition status was good while 153 (62%) participants had a bad nutrition status.

Linear regression results show the significance of .000, Coefficient of -61 and 153 as the Constant, implying that TB cases are likely to occur when exposed individuals have a bad nutrition status. The bar graph in Fig. 13 and the linear regression results in Table VIII summarizes the nutrition status of the participants.

e) Smoking: The participants who were smoking prior to being diagnosed with TB made up 44% which was 108 out of the 245 and those who never smoked were 137 which translates into 56% of the participants. The pie chart in Fig. 14 shows the comparison between the smoking and non-smoking participants.

f) Occupation: The study also sought to disaggregate the participants by occupation. The results reported 34 different occupation types for the 245 participants. There were 55 small scale traders who were selling in small shops and markets, 18 drivers who included bus, taxi or truck drivers, 9 bus conductors, 6 restaurant workers who worked as waiters or waitresses, 21 were school pupils, 18 who did general work and 52 were unemployed. Table IX shows selected occupation types of the participants.

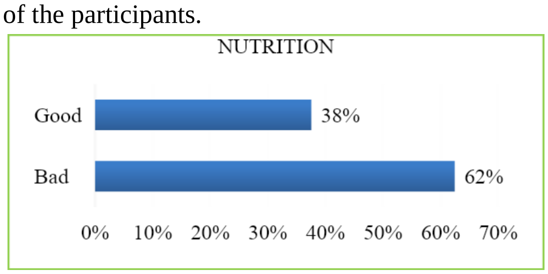


Fig. 13. Nutrition Status 12 Months Prior to TB Diagnosis.

TABLE. VIII. SMOKING VS TB CASES REGRESSION ANALYSIS COEFFICIENTS OUTPUT

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	***Std. Error***	***Beta***		
(Constant)	153	.000		.	.
Nutrition	-61	.000	-1.000	.	.

a. Dependent Variable: TB Cases

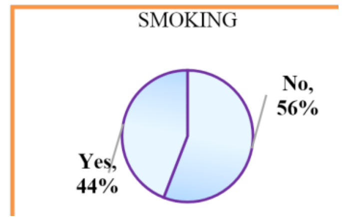


Fig. 14. Participants who Smoked Prior to TB Diagnosis.

TABLE. IX. SOME OCCUPATION TYPES OF PARTICIPANTS

Occupation	Number
Trader	55
Pupil	21
Bus conductor	9
Bus driver	8
Restaurant	6
Lorry/Truck driver	5
Taxi driver	4

The highlighted occupation types have one thing in common, they expose individuals to a lot of different people daily as they carry out their duties and mainly, in poorly ventilated spaces. This puts them at an increased risk of contracting TB compared to the people who have other occupation types.

B. *Prototype System*

Based on the proposed business model, the architectural system and the data designs, a prototype system has been developed. It is accessible for data entry on mobile devices (Smartphones, Tablets, Laptops, etc.) via a web browser. The system's front end was developed using PHP, embedded with the Google Mapping API for displaying geo distribution of TB cases in communities on the TB program managers' side. MySQL was used as the RDBMS.

a) Functional Requirements: Input: TB case details comprising parameters from which the likely cause of disease as well as geo coordinates of the location where the case was reported from.

System Processes: Data storage, retrieval, aggregation and summarization to produce maps and reports.

Output: Reports/Geo maps

b) Hardware Requirements: Basic Smartphones, Tablets or mobile computers can be used for data capturing as long as a web browser has been installed.

For good performance of the web application on the managers' side, a computer with the following hardware specification is recommended:

2.3 GHz processor speed or higher; 2GB of RAM or higher; 1GB of free disk space at least; and Internet connection.

c) Some Modules of the Prototype System: The system has two user types, the data capturers and program managers who need to make decisions to prevent the spread of TB. The data capturing module's primary function is to capture TB cases' data in communities and as such it only has a Data Entry button which allows the user to capture and send data to the server, and the Logout button for logging out of the system. Fig 15 on page 12 shows the data capturer's landing screen.

When logged in, the system takes note of the GPS location of the device and captures the geo coordinates of that position and shows the position of the device on the Google Mapping Application. This is for the purpose of recording the geo coordinates of the location of the TB case.

When the user clicks on the Add Entry button, the data capturing page loads onto the device. This is the page onto which the parameters of the TB case are recorded. These parameters are patient attributes and other factors which could be used to establish what could have caused the TB case in question and could be used when planning and implementing disease control and prevention activities.

It's important to note that the Location Access permission on the device being used for data capturing has to be On as shown in Fig. 16 when capturing data. This enables the application to access and pick the GPS location of the patient's place of residence.

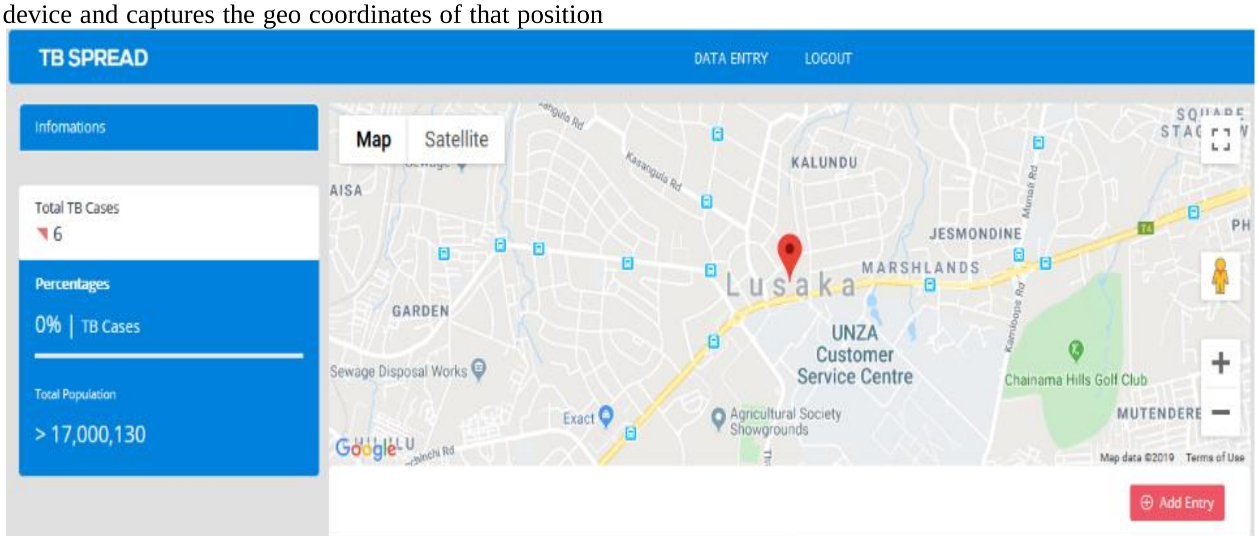


Fig. 15. Data Capturer Landing Page.

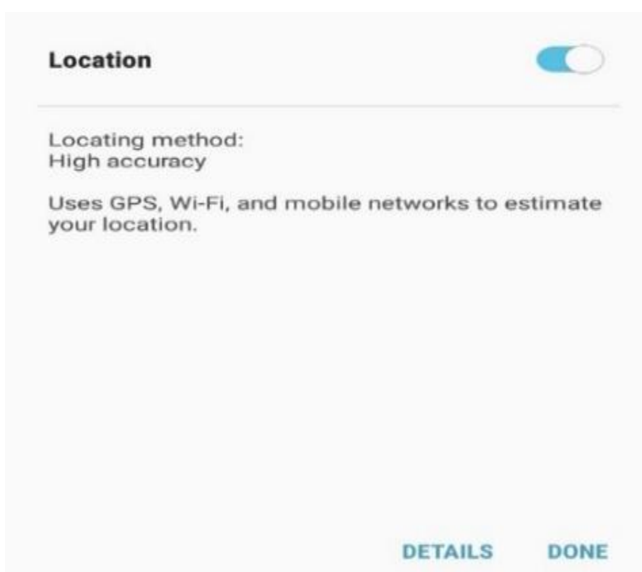


Fig. 16. Location Access Permission of the Data Capturing Device.

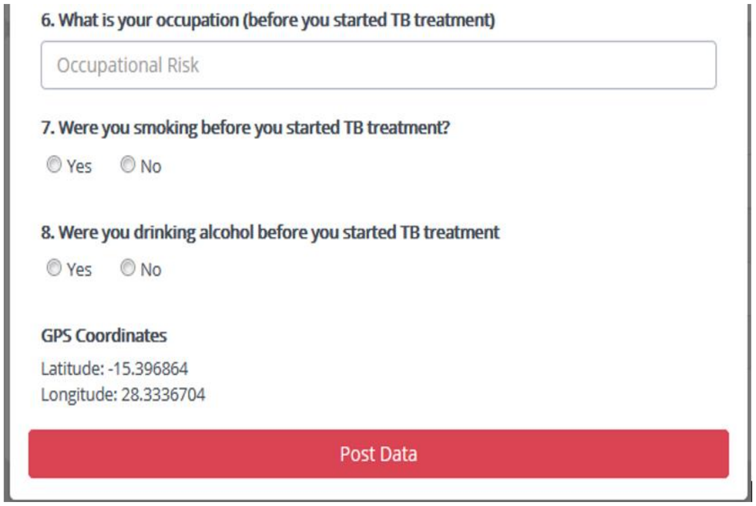


Fig. 17. Automatically Captured GPS Coordinates of the Data Entry Location.

The data capturer enters all the parameters of the TB case being reported except for the spatial data of the patient's location which is automatically picked based on the geo location of the device, which is the spatial location the TB patient. Fig. 17 is the screen shot of the data capturing page with automatically captured GPS coordinates of the location from which data is being captured.

To know the geographical distribution of TB cases for decision-making about the disease control and prevention activities which need to be carried out in a given area, TB program managers can view maps. The maps show the geographical location of TB cases, and these are likely targets when planning and implementing TB control and prevention activities. Fig. 18 shows the screen shot of some TB cases geographical distribution on a map

Additional information is availed to TB program managers through reports. The reports are accessible through the Reports button on the home page. The information on the reports can be used to craft interventions which can address the prominent causes of the disease in a given area. For example, if most patients where smoking prior to being diagnosed with TB, non-smoking messages can be prioritized when planning and implementing control and prevention activities. If the reports show that most patients were drinking prior to being diagnosed with TB, prevention activities must focus on drinking places as these could be breeding spaces for TB transmission. The screenshot in Fig. 19 on page 13 shows the draft report of the system prototype with summary patient parameters.

Fig. 18. Satellite Image Screen Shot of the Geographical TB Cases Distribution.

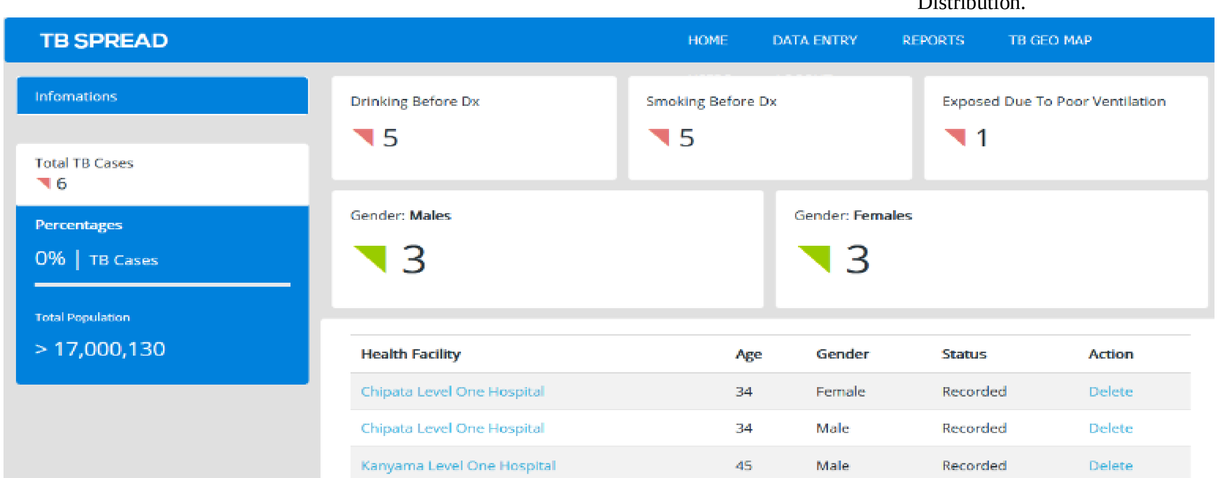


Fig. 19. Draft Report with Summary Patient Parameters.

VI. Discussion

The results of the data analysis show that crowding is the major driver of the TB transmission in Lusaka. While crowding is eminent, 67% (165 out of 245) of the participants' homes had poorly ventilated spaces which they shared with the people who were on TB treatment or presented with symptoms of the TB disease such as coughing and unexplained weight loss. From these findings, it can be established that crowding and poor ventilation resulted into prolonged exposure to the TB causing bacteria by individuals who contracted the disease. Exposure in the face of a weakened immune system and bad nutrition is likely to escalate the spread of TB. To reduce the transmission rates of the disease, prolonged exposure to the TB causing bacteria should be minimized and this can be achieved by improving ventilation systems of homes and other spaces where people are found. This can be achieved by having more openable windows installed in building such as houses, offices, bars, churches, schools and other public places and the windows should always be open whenever there are people. The baseline survey results show that crowding is a common and is one of the drivers of TB transmission and with the established poor ventilation patterns, TB is likely to keep spreading. To minimize the spreading of the disease, crowding must reduce. Another way of reducing the rate at which the disease develops, individuals must ensure they eat food with good nutrition value such as fruits, vegetables and food containing proteins as a good nutrition status is a deterrent to the development of active TB when an individual is exposed to the TB causing bacteria. Good nutrition status strengthens the immune system thereby reducing the susceptibility of individuals to the TB bacteria.

On the hand being male raises the likelihood of contracting TB to 65% compared to the 35% chances for the females. Analysis of the data showed that the average age of the participants was 33 years. Male participants made up 65% of the study population and 61% of the participants drank alcohol prior to TB diagnosis. These three variables, age, sex (males) and alcohol drinking coupled with a high rate of unemployment could be inferred to be driving factors in the spread of TB as men mostly spend time drinking alcohol in poorly ventilated and crowded drinking places such as taverns and bars. Equipping individuals who frequent such places with awareness knowledge of TB transmission and reducing on the times spent in such spaces can reduce the rate at which the disease is spread.

From the prototype development perspective, it can be established that when properly harnessed, ICTs can be used to capture, process information and provide it in way and formats which are easy for decision maker to make decisions.

Though there could be other factors which promote the spread of TB, TB program managers can put in interventions which can help to reduce the rate at which the disease is spread if they have this information and other reports readily availed by the system. Coupling this information with the geospatial display of TB cases on a mapping application can help TB program managers to put resources where it matters the most in order to fight TB.

VII. Conclusions and Recommendations

A. *Conclusion And Recommendation*

This paper presents a way of making well informed decisions in the face of constrained resources to help with the fight against TB. It utilizes the flexibility of cloud computing, the accuracy and reliability of geo spatial analysis and display of information on a mapping application made feasible by GIS and the ability to create reports and help to make decisions via the web application. It is recommended that information is always used as a cornerstone for decision-making in TB prevention and this paper makes this recommendation a reality.

B. *Future Works*

While this paper focused on the use of ICTs to help TB program managers make decisions based on the information availed to them by the system, it is desirable in future, to have the system it self suggest the best intervention activities to be carried out in a given community to the TB program managers. Having the system recommend the intervention packages to be implemented will likely result in more efficient and impactful interventions because the system will be making decisions and not humans, thereby reducing the inherent human biasness in the control and prevention activities planned and implemented.

The Reports module need to be enhanced and expanded to cover more parameters upon which TB program managers can base decisions when planning and implementing disease control and prevention activities. It also needs to be visually improved to quicken decision-making.

AcknowledgEment

We are very grateful to everyone who contributed towards the development of this paper, especially to the supervisors for the guidance, knowledge and encouragements, and to the National TB Program Manager for the encouragement and the support rendered when we were seeking permission from the Ministry of Health to conduct the baseline survey.

References

[1] World Health Organization, Global tuberculosis report 2019, World Health Organization, Geneva, Switzerland, 2019.

[2] P.J. Dodd, C. M. Yuen, C. Sismanidis, J. A. Seddon, H.E. Jenkins, "The global burden of tuberculosis mortality in children: a mathematical modelling study," Lancet Glob Health Vol 5 September 2017, pp. e898–906.

[3] G. Churchyard. et al, 2017, "What we know about tuberculosis transmission: an overview," The Journal of Infectious Diseases, JID 2017:216 (Suppl 6), pp. S629-s635.

[4] World Health Organization, "Global tuberculosis report 2018", World Health Organization, Geneva, Switzerland, 2018.

[5] Government of the Republic of Zambia - Ministry of Health, National Strategic Plan (NSP) for Tuberculosis Prevention, Care and Control (2017-2021) - Towards Elimination, Lusaka, Zambia, 2016.

[6] Mutembo et al, 2019, "Urban-rural disparities in treatment outcomes among recurrent TB cases in Southern Province, Zambia," BMC Infectious Diseases (2019) 19:1087.

[7] UNAIDS, UNAIDS DATA 2018, Joint United Nations Programme on HIV/AIDS ,Geneva,Switzerland, 2018.

[8] I. Mwila and J. Phiri, "Geospatial, cloud and web based model for evidence-based decision-making in tuberculosis prevention," in

International Conference in ICT (ICICT 2019) – Lusaka, Zambia (20th – 21st November 2019), ISBN 978-9982-70-985-9, Page 110 – 115.

[9] A. Ali. 2011. "Exploring the aspects of digital divide in a developing country," Issues in Informing Science and Information Technology Volume 8, 2011.

[10] A. Joshi, et al, 2013, "The role of information and communication technology in community outreach, academic and research collaboration, and education and support services (IT-CARES)," Perspectives in Health Information Management, Fall 2013.

[11] J. Doosje, A. Bosman, M. L. A. Heijnen, V. van Kuyvenhoven, A. A. Warris and J. van Steenbergen, 2004, "Improving the use of information and communication technology in infectious disease control," European Journal Of Public Health Vol. 14 2004 NO. 4.

[12] A. Rukshan, A. Miroshan, "An ICT-based real-time surveillance system for controlling dengue epidemic in sri lanka," unpublished.

[13] World Health Organization. Digital Health For The End Tb Strategy: An Agenda For Action, Geneva, Switzerland, 2015.

[14] W. Curioso, et al, Information and communication technologies for prevention and control of HIV infection and other STIs, Washington, United States, 2017.

[15] B. C. K. Choi, 2012, "The past, the present and the future of public health surveillance," Scientifica, Volume 2012, Article ID 875253.

[16] P. Narasimhan, J. Wood, C. R. MacIntyre and D. Matai, 2013. "Risk factors for Tuberculosis," Pulmonary Medicine Volume 2013, Article ID 828939.

[17] B. Mathema, 2017, "Drivers of Tuberculosis Transmission," The Journal of Infectious Diseases, JID 2017:216 (Suppl 6), pp. S645-s653.

[18] P. Wargocki, 2013, "The Effects of Ventilation in Homes on Health," International Journal of Ventilation ISSN 1473-3315 Volume 12 No 2 September 2013.

[19] M. Ruiz-Castell, et al, 2015. "Household crowding and food insecurity among inuit families with school-aged children in the Canadian Arctic," American Journal of Public Health| March 2015, Vol 105, No. 3.

[20] S. Kevin, R. L. Blake. K. A. Simic, Measuring Overcrowding in Housing Prepared for: U.S. Department of Housing and Urban Development Office of Policy Development and Research, Econometrica, Inc. Bethesda, Maryland, United States, 2007.

[21] J. Doosje, et al, 2004. "Improving the use of information and communication technology in infectious disease control," European Journal Of Public Health Vol. 14 2004 No. 4.

[22] United States Centre for Disease Control and Prevention (CDC), Know Your Epidemic: Disease Surveillance For Targeting Efforts To End TB, 2015.

[23] N. Sabi'u, S. N. Muhammed, N. Zakari and M. S. Khalil, 2015, "Vector data model in GIS and how it underpins a range of widely used spatial analysis techniques," Dutse Journal of Pure and Applied Sciences 1(1) June 2015 pp. 122 - 132.

[24] Rezaeian, et al, 2007. "Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary," J Epidemiol Community Health 2007; 61:98–102.

[25] R, Zerbe, et al, An Analysis of Benefits from use of Geographic Information Systems By King County Washington, Richard Zerbe and Associates, Seattle WA, 2012.

[26] K. Soniya, A. S. Kumar. "A basic study on cloud computing," IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, pp 43-46.

[27] S. Chowdhary and A. Srivastava, 2013, "Cloud computing: a key to effective & efficient disease surveillance system," in Proc. of Int. Conf. on Advances in Computer Science and Application 2013.

[28] Veer Surendra Sai University of Technology. Internet & Web Technology-Ii Lecture Notes. Department of Computer Science & Engineering, IT.

[29] G. Pullaiah College of Engineering and Technology. Lecture Notes on Web and Internet Technologies. Department of Computer Science and Engineering.

[30] J. Choi, et al, 2016, "Web-based infectious disease surveillance systems and public health perspectives: a systematic review," BMC Public Health (2016) 16:1238.

[31] J. J. Hamilton and R. S. Hopkins, The CDC Field Epidemiology Manaul - Using Technologies for Data Collection and Management, 2018.

[32] R. Islam, et al, 2010, "Mobile application and its global impact," International Journal of Engineering & Technology IJET-IJENS Vol: 10 No: 06.

[33] Y. P. Li, et al, 2013, "Decision support system for the response to infectious disease emergencies based on webgis and mobile services in China," PLoS ONE 8(1): e54842. doi:10.1371/journal.pone.0054842.

[34] E. Almrot and S. Andersson, 2013, "A study of the advantages & disadvantages of mobile cloud computing versus native environment," unpublished.

[35] R. Chunara, 2012, "New technologies for reporting real-time emergent infections," Parasitology, 2012 December; 139(14): 1843–1851. Doi: 10.1017/S0031182012000923.

[36] N. Beyers, et al, 1996, "The use of a geographical information system (GIS) to evaluate the distribution of tuberculosis in a high incidence community," Oepariments of Paediatrics and Child Health and Medical Biochemistry, and Institute for Geographical Analysis, Volume 86 No.1 January 1996 SAMJ.

[37] N. Tiwari, et al, 2006, "Investigation of geo-spatial hotspots for the occurrence of tuberculosis in Almora district, India, using GIS and spatial scan statistic," International Journal of Health Geographics 2006, 5:33 doi: 10.1186/1476-072X-5-33.

[38] B.L. Gleason1, A. Kamara, N. Clemens and D. Kargbo, 2018, "Establishment of an electronic integrated disease surveillance and response system in Sierra Leone," International Journal of Infectious Diseases 79(S1) (2019) 1–150.

[39] S. Biswas, Anisuzzaman, T. Akhter, M. S. Kaiser and S. A. Mamun, 2016, "Cloud based healthcare application architecture and electronic medical record mining: an integrated approach to improve healthcare system," Institute of Information Technology Jahangirnagar University, Dhaka, Bangladesh.

[40] M. O. Lwin, et al, 2017, "Flumob: enabling surveillance of acute respiratory infections in health-care workers via mobile phones," Front Public Health 5:49. doi: 10.3389/fpubh.2017.00049.

Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List

Gayatri Venugopal-Wairagade[1], Jatinderkumar R. Saini[2], Dhanya Pramod[3]
Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India[1, 2]
Symbiosis Centre for Information Technology, Symbiosis International (Deemed University), Pune, India[3]

***Abstract*—This paper is an effort to complement the contributions made by researchers working toward the inclusion of non-English languages in natural language processing studies. Two novel Hindi language resources have been created and released for public consumption. The first resource is a corpus consisting of nearly thousand pre-processed fictional and non-fictional texts spanning over hundred years. The second resource is an exhaustive list of stop lemmas created from 12 corpora across multiple domains, consisting of over 13 million words, from which more than 200,000 lemmas were generated, and 11 publicly available stop word lists comprising over 1000 words, from which nearly 400 unique lemmas were generated. This research lays emphasis on the use of stop lemmas instead of stop words owing to the presence of various, but not all morphological forms of a word in stop word lists, as opposed to the presence of only the root form of the word, from which variations could be derived if required. It was also observed that stop lemmas were more consistent across multiple sources as compared to stop words. In order to generate a stop lemma list, the parts of speech of the lemmas were investigated but rejected as it was found that there was no significant correlation between the rank of a word in the frequency list and its part of speech. The stop lemma list was assessed using a comparative method. A formal evaluation method is suggested as future work arising from this study.**

***Keywords*—Hindi; corpus; aesthetics; stopwords; stoplemmas**

I. Introduction

One of the basic requirements to devise a tool to perform any task in Natural Language Processing (NLP) is a corpus that represents the target language or the target domain. Adhering to the ‘Bender Rule’, according to which researchers are required to name the language that was targeted by the study, we would like to inform the readers that the study focuses on Hindi, the official language of India. The language ranks third on the list of the languages with the largest number of first language speakers in the world [1]. The study aims at building a corpus in the aesthetics domain and utilizing the corpus, along with other corpora to publish an exhaustive list of stop lemmas based on their raw frequencies in the corpora. The corpus has been released under the GNU General Public License for public use, in order to ease the process of corpus acquisition. It is also necessary to mention that the text and the corpora that were acquired from various sources have been used solely for academic and research purpose. The corpus was created because of the difficulty that we faced while searching for and acquiring novels, stories and non-fictional content written by contemporary authors as well as content written by authors in India’s pre-independence era, i.e., prior to 1947. The broad objective of the study is to utilize and release unbiased time-independent text in the form of a corpus. Since the study emphasizes the lexical aspect of text, features related to context and discourse have not been discussed in this paper. The stop lemma list created as part of this work has been built using text from multiple sources, thus making it suitable for generic consumption. The reported outcome is best as on date subject to the data used for this research. We believe that this corpus and the list of stop lemmas would be a useful resource for NLP tasks such as creating language models, text classification and information retrieval in Hindi.

The remaining paper is organized into three sections. Section II of the paper consists of the existing work in this area and the research questions. Section III contains the description of the corpus and the methodology along with the final list of stop lemmas. Section IV consists of the concluding remarks.

II. Existing Work and Objectives

Stop words are words that are present in a sentence solely for grammatical reasons and do not contribute to the information obtained from the text [2]. Hence if these stop words are identified and removed before using the text for a task, the performance of the task could improve. Many studies have focused on the importance of removing stop words as a pre-processing step for text processing tasks [3], [4] and [5]. Author in [6] manually extracted stop words based on parts of speech such as pronouns, prepositions, conjunctions etc. from two news-based corpora to create a stop word list in Hindi consisting of 275 words. Author in [7] created a stop word list by converting words to lemmas in a corpus of news articles consisting of 441,153 words. Author in [8] proposed a method for automatic stop word generation that created stop word lists that matched the top twenty stop word lists from four publicly available lists. Their corpus was based on news articles. Although we found studies that focused on automatic stop word generation [9], [10], [11], [12], [13] and others that published the stop word lists for public use [14], we could not find an exhaustive publicly available list of stop words in Hindi based on multiple corpora. Another problem with multiple lists based on one or two corpora is the inconsistency of the words in the lists [15]. Author in [16] manually created a stop word list consisting of 256 words from Punjabi poetry and articles. This list was brought down to 184 unique words by lemmatizing the words. We would like to emphasize here that the researchers lemmatized the words after identifying a word as a stop word as opposed to lemmatizing all the words and then identifying the stop words. [17] collated a list of stop words in numerous languages spanning multiple countries.

The researchers observed the presence of multiple word forms in the stop word lists depending on the morphological complexity of the language. We catered to this issue while framing the objectives of our study. Author in [18] manually created a list of stop words based on news corpora in Gujarati. The researchers also aimed to find a pattern in the stop words that were assigned the same part of speech tag but they could not find any similarity. Another study used a threshold based approach to identify stop words in Sanskrit text by calculating the frequencies of the words in the text [19]. The researchers removed nouns from the list under the assumption that nouns cannot be treated as stop words. We aimed to test the validity of this assumption through our study.

The source of data in majority of the studies discussed in this section was news articles. All these studies, except for one, focused on the creation of stop word lists, instead of stop lemma lists. This would lead to the creation of a list that consists of a morphological variation of a word instead of the root word itself, which could negatively affect the quality of the NLP task owing to the non-robust nature of the list, hence defeating the purpose of its creation. One of the studies focused on assigning a part of speech tag depending on the stop word. We formed one of our objectives along similar lines, but our emphasis was on the relationship between parts of speech and stop words in general, rather than on categorizing similar stop words under a particular lexical category.

The objectives of the study were framed as follows, based on the analysis of the existing work:

a) To determine whether the existing publicly available stop word lists are inter-changeable

b) To determine whether the part of speech of a word can be used to determine whether the word is a stop word or not

c) To create an exhaustive stop lemma list irrespective of the target domain

In order to meet the above objectives, our study was focused on finding answers to the following research questions:

RQ1: Are the top ten stop words across all the available stop words lists the same, irrespective of the rank of the words in the lists?

RQ2: Is there any significant relationship between:

a) The part of speech of a word and its rank in the stop word list?

b) The part of speech of a lemma and its rank in the stop lemma list?

RQ3: How can the available resources be synthesized to create an exhaustive stop lemma list?

III. Methodology

A. Corpus Creation and Corpora Metadata

This section discusses the creation of our corpus and throws light on the metadata of the corpora we collected. We scraped novels and stories from http://hindisamay.com, an e-library maintained by Mahatma Gandhi Antarrashtriya Hindi Vishwa Vidyalaya (translated to Mahatma Gandhi International Hindi University), Wardha, http://premchand.co.in, a website dedicated to the popular novelist Premchand's stories, and Bhandarkar Oriental Research Institute's Digital Library (http://borilib.com). Scrapy, an open source tool, was used to extract content from websites. We also extracted content from PDF of novels that were not available on the specified websites, but the text thus extracted could not be used because of encoding errors. Owing to issues such as this and lack of availability of public content, the size of the corpus is not comparable with that of articles present in Wiki dumps. However, the corpus, in combination with other corpora can become an effective resource.

As a preprocessing step, we split the text into sentences, tokenized the sentences and deleted special characters, English tokens and Latin numbers. Joined words were not segmented. The details with respect to the vocabulary of our corpus and the corpora that we acquired from various sources to create a stop word list, can be seen in Table I. LR indicates the language resource specified in the Language Resources section. LR#12 refers to the corpus that we created.

The numbers presented in the table against LR#13 to LR#14 were generated by calculating the raw frequency of the words, and by fetching the lemmas of the words present in the corpora and calculating the raw frequency of the lemmas, as this information was not readily available. The corpora included in the table span over ten domains and a period of over hundred years.

The state-wise distribution of the authors whose works have been included in the corpus, can be seen in Fig. 1.

As depicted in the figure, majority of the work is associated with authors whose native state is Uttar Pradesh, a state in northern India. Another observation made in the metadata was that only 4.84% of the publications are authored by females. This number was 0 in the pre-independence era for this corpus. However, the number increased as the years passed, although it is incomparable with that of the male authors.

TABLE. I. Metadata of the Corpora that are Part of the Study

S.No.	Source	Unique Word Count	Unique Lemma Count	Domain
1	LR#12	145508	118266	Aesthetics
2	LR#13	21335	17159	Entertainment
3	LR#14	119313	102201	Not available
4	LR#15	2330	1851	Varied domains
5	LR#16	21826	18220	Tourism
6	LR#17	39351	32074	Agriculture and Entertainment
7	LR#18	35018	28645	Agriculture, Entertainment, Politics and Public Administration, Sports, Religion, Literature, Aesthetics, Economy
8	LR#19	20430	16673	Health

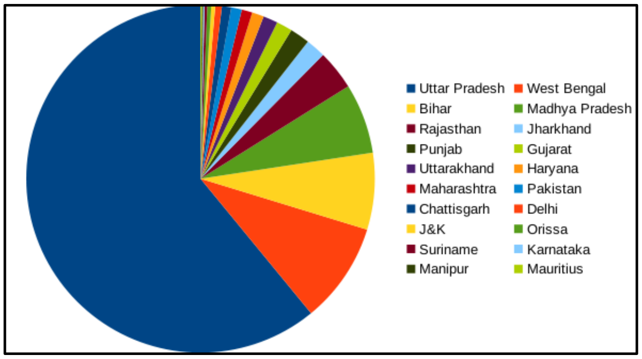


Fig. 1. State-Wise Distribution of Authors in the Corpus.

978 articles including novels, short stories and non-fictional texts were collated from the sources mentioned earlier. Out of these 978 articles, the metadata of 164 articles could not be found. We were unable to find details of authors who are not highly accomplished, but we believe that the authors are not amateur writers since the university chose to publish their work on their website.

B. Stop Word List Creation

In this section, we attempt to answer the research questions framed in Section 2.

RQ1: Are the top ten stop words across all the available stop words lists the same, irrespective of the rank of the words in the lists?

The null hypothesis was framed as follows:

H_0: The top ten stop words across all the available stop words lists are the same, irrespective of the rank of the words in the lists.

H_A: The top ten stop words across all the available stop word lists are not the same, irrespective of the rank of the words in the lists.

We created a list of the top ten stop words obtained from publicly available lists and generated the list from corpora released by Technology Development for Indian Languages (TDIL), Ministry of Electronics & Information Technology (MeitY), Government of India., Centre for Indian Language Technology (CFILT), IIT Bombay, Wikipedia dump, http://opus.nlpl.eu that consists of subtitles obtained from opensubtitles.org, and the aesthetics corpus that we created using articles collected from the sources mentioned above. We did not have access to the corpora that were used to generate the publically available lists as these lists have neither been released as formal resources nor has the process of creation of these lists been published as research papers. Table II contains the list of the top ten stop words in each source.

A word cloud generated from the list of these stop words and their count in the lists across all the sources, can be seen in Fig. 2. The size of the word in the figure is directly proportional to its count. It was observed that the word lists were not consistent across the sources. Out of the 19 sources and 82 unique stop words, the maximum count of sources in which a word appeared was 14.

TABLE. II. TOP TEN STOP WORDS IN EACH SOURCE

Source	Stop Words									
LR#1	में	है	हैं	नहीं	लिए	गया	तथा	अपने	कुछ	साथ
LR#2	जैसा	मैं	उसके	कि	वह	था	के लिए	पर	हैं	साथ
LR#3	की	और	एक	तक	में	है	आप	कि	यह	वह
LR#4	के	है	में	की	से	और	का	को	हैं	पर
LR#5	एक	आप	और	यह	कर	हम	वह	पर	इस	अब
LR#6	अत	अपना	अपनी	अपने	अभी	अंदर	आदि	आप	इत्यादि	इन
LR#7	अंदर	अत	अदि	अप	अपना	अपनि	अपनी	अपने	अभि	अभी
LR#8	पर	इन	वह	यिह	वुह	जिन्हें	जिन्हों	तिन्हें	तिन्हों	किन्हों
LR#9	मैं	मुझको	मेरा	तुम्हारा	हमने	हमारा	अपना	हम	आप	आपका
LR#10	के	का	एक	में	की	है	यह	और	से	हैं
LR#11	और	पर	एक	रत	कर	इस	यह	अन	वर	सम
LR#12	है	के	में	की	और	से	का	को	नहीं	तो
LR#13	है	के	मैं	नहीं	में	हैं	एक	आप	और	लिए
LR#14	के	है	में	की	से	और	का	को	हैं	पर
LR#15	के	में	की	और	लिए	हैं	है	से	का	को
LR#16	के	है	में	और	की	से	एक	का	हैं	को
LR#17	के	है	में	की	का	से	और	को	हैं	भी
LR#18	के	है	में	की	का	से	को	और	हैं	ने
LR#19	है	के	में	से	की	को	का	हैं	और	हो

Fig. 2. Word Cloud based on the Count of the Top Ten Stop Words Across 19 Sources.

We also found inconsistencies in the list of lemmas. The list of top ten stop lemmas from each source can be seen in Table III. Table III consists of fewer language resources as compared to Table II, as there exists to publicly available lists of stop lemmas at the time of writing this paper. Hence, we used the corpora in hand to generate the lemmas of the words and sorted them based on their raw frequencies.

A word cloud generated from the list of the top 10 lemmas and their count in this list across 8 sources (LR#12 to LR#19), can be seen in Fig. 3.

An interesting observation here was that among 8 sources and 22 unique stop lemmas, the maximum count of sources in which a lemma appeared was 8, which is a significant improvement over the count of the words in the stop word list. Hence, though we reject our null hypothesis and accept the alternative hypothesis, we believe that while creating an exhaustive list of stop words, we should consider the lemmas of the words and not just various morphological variations of a word.

RQ2: Is there any significant relationship between:

a) The part of speech of a word and its rank in the stop word list?

b) The part of speech of a lemma and its rank in the stop lemma list?

TABLE. III. TOP TEN STOP LEMMAS IN EACH SOURCE

Source	**Stop Lemmas**								
LR#12	का	है	वह	हो	में	कर	था	जा	यह	और
LR#13	है	का	मैं	कर	वह	यह	नहीं	हो	में	आप
LR#14	का	है	में	वह	हो	कर	यह	जा	से	और
LR#15	का	है	में	और	कर	जा	से	को	दे	यह
LR#16	का	है	कर	में	और	से	एक	जा	यह	ले
LR#17	का	है	कर	में	यह	हो	जा	से	और	वह
LR#18	का	है	कर	में	यह	हो	वह	से	को	जा
LR#19	है	का	कर	में	हो	से	यह	जा	को	और

Fig. 3. Word Cloud based on the Count of the Top Ten Stop Lemmas Across 8 Sources.

The null hypotheses were framed based on a general assumption that parts of speech such as prepositions, postpositions and symbols are commonly marked as stop words.

a) H_0: There is a significant relationship between the part of speech of a word and its rank in the stop word list.

H_A: There is no significant relationship between the part of speech of a word and its rank in the stop word list.

b) H_0: There is a significant relationship between the part of speech of a word and its rank in the stop word list.

H_A: There is no significant relationship between the part of speech of a lemma and its rank in the stop lemma list.

In order to test the hypotheses, we considered the parts of speech of all the words that appeared in the top ten ranks of the stop word lists for each source from S12 to S19. The parts of speech for every word and every lemma in each list was determined for all the sources using the part of speech generated from the Hindi treebank of the stanfordnlp POS processor. The point biserial correlation coefficient between the parts of speech present in the top ten ranks of the stop words lists, and the rank of the word in the list was calculated. The parts of speech of 10722288 unique words and 335089 lemmas were analyzed. The results can be seen in Table IV. The table consists of descriptive statistics of the coefficient values across the lists generated from all the sources.

It was observed that there was no significant relationship between any part of speech and the words' ranks. Hence the null hypotheses were rejected and the alternative hypotheses were accepted. This bursts the assumption that words belonging to parts of speech such as prepositions, postpositions, conjunctions etc. would be ranked among the top stop words.

RQ3: How can the available resources be synthesized to create an exhaustive stop lemma list?

TABLE. IV. DESCRIPTIVE STATISTICS OF THE CORRELATION BETWEEN THE PART OF SPEECH OF THE WORDS AND THEIR CORRESPONDING RANK IN THE LIST OF STOP WORDS

Part of Speech	**Source List**	**Point Biserial Coefficient Value**				**P-Value**	
		Mean	***SD***	***Max***	***Min***	***Mean***	***SD***
NN/NNP/ NNPC	Words	-0.02	0.09	0.09	-0.18	0.05	0.10
	Lemmas	-0.02	0.11	0.22	-0.16	0.04	0.11
PSP/PRP	Words	-0.07	0.04	-0.01	-0.11	0.06	0.16
	Lemmas	-0.06	0.04	-0.01	-0.11	0.06	0.16
SYM	Words	0.10	0.06	0.16	0.00	0.10	0.31
	Lemmas	0.11	0.06	0.20	0.04	0.07	0.18
VM	Words	-0.04	0.06	0.09	-0.08	0.13	0.27
	Lemmas	-0.03	0.06	0.09	-0.08	0.20	0.29
QC/QF/QO	Words	-0.05	0.02	-0.02	-0.08	0.05	0.12
	Lemmas	-0.06	0.03	-0.02	-0.10	0.04	0.13
NEG	Words	-0.02	0.02	0.00	-0.05	0.08	0.05
	Lemmas	-0.02	0.02	0.00	-0.06	0.15	0.22
CC	Words	-0.04	0.02	-0.01	-0.07	0.08	0.24
	Lemmas	-0.04	0.02	-0.01	-0.07	0.08	0.22

In an attempt to solve this question, we collected the top 100 stop words from the existing publicly available lists, took a union of all the lists, and replaced the words with their lemmas. The lemmas were used owing to the observation made while searching for the answer to RQ1. This was named Set A. Set B was created by taking a union of the top 100 stop lemmas from all the corpora. The exhaustive stop word list C was another set that was generated by taking an intersection of Set A and Set B.

Set A = genLemma(Set_1) U genLemma(Set_2) U genLemma(Set_n),

where Set_i represents the set of stop words in source i; 1<=i<=11 and genLemma is a function that replaces each word in the set with its corresponding lemma

Set B = Set_1 U Set_2 U ... Set_m,

where Set_j represents the set of the top 100 most frequent lemmas in source j; 12<=j<=19

Set C = Set A ∩ Set B

1096 words from publicly available lists were brought down to 1071 words after removing the duplicate words and phrases. A list of unique lemmas of these words was created. The length of this list was 370.

Set B, which was created by fetching the lemmas of all the words in the corpora, consisted of 213554 lemmas. These lemmas were generated from 13811781 words that were brought down to 405111 words after removing the duplicates.

The final list of stop lemmas was created by combining both the lists and extracting the common lemmas. The list consists of 311 lemmas, as can be seen in Table V. The lemmas have been arranged in decreasing order of frequency from top to bottom.

The complete list is available on https://github.com /gayatrivenugopal/hindi-corpus-stoplemmas.

In order to assess this list, we compared our list with the English stop word list present in Python's Natural Language ToolKit package (nltk). We used Google Translate with manual intervention to translate the English stop words into their equivalent Hindi words. We chose English as we could not ascertain the correctness of translation of other languages in Hindi. The lemmas of these translated words in Hindi were generated and compared with the exhaustive stop lemma list. 74 unique equivalent Hindi stop lemmas were generated for the 179 words in the English stop word list. We could not translate the words 'being', 'will' and 'shall'. Although 'shall' and 'will' were not present as it is in the English stop word list, we expanded the word form ''ll' to 'shall' and 'will'. An ambiguity of a particular kind was found while analyzing the words in the English stop words list. The list contained the word 'won', which was next to words such as 'shan' and 'aren'. Here, we did not consider the meaning of 'won' to be the past tense of 'win'. On the other hand, it was a form of wouldn't, i.e., won't, wherein the apostrophe and the 't' were removed from the word. We used the same approach for words such as 'shant', 'aren', etc. Following the disambiguation, translation and lemmatization phase, we observed that out of the 74 unique lemmas, 73 lemmas were present in our stop lemma list. The lemma that was absent from our list was 'जरूर', the Hindi equivalent of the English stop word 'must'.

TABLE. V. 311 STOP LEMMAS PRESENT IN THE EXHAUSTIVE LIST

का	लेकिन	बाहर	अथवा	दर	तर	टर
है	तरह	पूछ	मत	गलत	कौनसा	यक
वह	अब	भारत	पुरुष	खिलाफ	कोन	मह
में	बहुत	छोटा	रास्ता	जन	लत	पनी
कर	दिन	सामने	जैसे	पालन	निहायत	उनकि
हो	रख	बीच	आवश्यकता	आठ	कवर	तथ
यह	जब	तीन	कल	जोड़	बंदरगाह	उनक
जा	लगा	हर	भीतर	पल	दुसरा	उत
और	तथा	जहाँ	कोशिश	बिलकुल	तिन्ह	वग़ैरह
से	बाद	केवल	प्रत्येक	उच्च	बाला	रक
था	चाह	डाल	औरत	सदा	तिस	वुह
को	यहाँ	कितना	दस	अधिकांश	उह	
मैं	दूसरा	बना	खाना	वन	तिसे	
नहीं	घर	वर्ष	दौरान	निकट	यत	
पर	समझ	रात	वहीं	छह	एलन	
रह	चाहिए	सबसे	सुबह	आर	किर	
भी	रूप	कैसे	वर्ग	ओ	जर	
कि	जैसा	माँ	डर	पृथ्वी	डल	
तो	पहले	बारे	पढ	बज	पडा	
ले	बार	आगे	ए	हल	रत	
एक	कभी	भाग	तुम्हारा	छत	अल	
दे	अच्छा	पी	मा	थक	गर	
ही	बोल	अलग	सच	दोपहर	चकमक	
ने	कारण	कहाँ	मतलब	तहत	उम	
अपना	ओर	विकास	मानो	पे	चन	
जो	हाथ	प्राप्त	माध्यम	ओह	दक	
आ	कौन	कार्य	अंत	तस्वीर	नक	
कह	आज	जगह	उधर	जादू	सकत	
कोई	पास	ऊपर	कब	बिंदु	बर	
हम	पूरा	शब्द	कुल	सन	आत	
सक	वहाँ	बस	जबकि	शक	आद	
आप	अधिक	ज्यादा	संख्या	असल	ईस	
कुछ	भर	भाषा	एस	धन्यवाद	तया	
देख	सुन	बदल	ना	एल	खक	
बात	द्वारा	नीचे	हवा	आह	सबस	

साथ	देश	बंद	परिवर्तन	लय	रण	
क्या	बता	मर	सर	सहमत	मक	
दो	क्यों	काफी	बहन	पहल	करत	
तक	सारा	खुद	सोना	नफरत	यन	
ऐसा	प्रकार	बड़ा	वजह	मुझको	उद	
लग	बडा	पिता	मात्र	नरक	साबुत	
चल	नया	शहर	मदद	दुबारा	अर	
सब	निकल	दुनिया	प्रकाश	गत	यर	
बन	लिख	विशेष	खबर	सेट	लन	
लोग	पानी	जितना	आग	जनरल	षण	
मिल	इसलिए	उपयोग	पद	वर	एसे	
या	एवं	अंदर	अत	अप	क्यूंकि	
फिर	कई	खेल	बह	दोनो	पत	
वाला	अभी	लगभग	आराम	नह	वत	
लिए	अगर	स्वयं	रस	आध	रद	

IV. Conclusion

At the onset, we set out to create a corpus consisting of aesthetic texts in Hindi and to study the stop words present in this corpus as well as other existing corpora and publicly available lists, with the intent of finding a comprehensive list of stop words. Through this study, we have achieved the following broad objectives:

a) To create an unbiased time-independent aesthetics corpus in Hindi

b) To study the relationship between the part of speech of a word and its rank in the frequency list

c) To provide an exhaustive stop lemma list in Hindi

In order to achieve the first objective, we used texts belonging to fictional and non-fictional categories.

An aesthetics corpus consisting of approximately 1000 articles and 145508 words was built using text from various sources. The gender distribution of the authors was discussed. We could not find significant amount of work by female authors that is available in digital format for public consumption. The corpus consists of both, contemporary texts as well as stories that are over a hundred years old. In order to create an exhaustive stop word list, texts were collated from seven different corpora. The stop word lists that are available for public use were also used. Stop word lists were generated from each corpus based on the word count. It was found that there was an inconsistency even in the top ten stop words across all the lists. We also studied the individual lists and corpora to determine whether the part of speech of a word could be used to determine whether the word can be considered to be a stop word but the results contradicted our assumption. Owing to this observation, while creating a comprehensive list, we did not focus on the part of speech as a characteristic of a word to deem it as a stop word. Only the frequency of the word was taken into count. Through our study, we propose the concept of 'stop lemma' list, instead of mere 'stop word' list, as it is believed to yield better results when used in NLP tasks. Therefore in order to make the list robust, the lemmas of the words were considered as it was found that stop lemmas were more consistent than stop words across different corpora, as can be observed from the density of the respective word clouds generated using the top ten words and the top ten lemmas in the sources . At the time of writing this paper, such a list does not exist in the public domain. This list is limited to the texts collected from the specified sources. We assessed the quality of the list by translating, lemmatizing and comparing the English stop words present in NLTK (a Python package that is widely used in English NLP tasks), with the lemmas from our list. We chose English owing to the familiarity with the language as it helped in determining the quality of translation. We found that almost every lemma that was present in the English list mapped to a corresponding lemma in the Hindi list. Future work arising from this study could comprise evaluation of the list by using it with different NLP tasks, as well as a comparison of the results of a specific NLP task with respect to the use of stop words and stop lemmas in various languages. The corpus, the metadata as well as the analysis has been made available for public consumption at https://github.com/gayatrivenugopal/hindi-corpus-stoplemmas.

Acknowledgment

This study is funded by Symbiosis Centre for Research and Innovation, Symbiosis International (Deemed University). We would also like to thank the organizations that provided the corpora that were used in the study, and the Stanford NLP group that released the stanfordnlp package for Python which was used to fetch accurate lemmas of the words in Hindi. The "Linguistic Resources" obtained from TDIL have been developed & made available by TDIL, MeitY, Government of India. The resources from CFILT, IIT-Bombay have been downloaded from http://www.cfilt.iitb.ac.in/Downloads.html.

References

[1] Ethnologue, https://www.ethnologue.com (2019). Accessed October 21 2019.

[2] Joshi, H., Pareek, J., Patel, R., & Chauhan, K. (2012, December). To stop or not to stop—Experiments on stopword elimination for information retrieval of Gujarati text documents. In 2012 Nirma University International Conference on Engineering (NuiCONE) (pp. 1-4). IEEE.

[3] Jha, V., Manjunath, N., Shenoy, P. D., & Venugopal, K. R. (2016, January). Hsra: Hindi stopword removal algorithm. In 2016 International Conference on Microelectronics, Computing and Communications (MicroCom) (pp. 1-5). IEEE.

[4] Silva, C., & Ribeiro, B. (2003, July). The importance of stop word removal on recall values in text categorization. In Proceedings of the International Joint Conference on Neural Networks, 2003. (Vol. 3, pp. 1661-1666). IEEE.

[5] Na, D., & Xu, C. (2015). Automatically generation and evaluation of stop words list for Chinese patents. Telkomnika, 13(4), 1414.

[6] Larkey, L. S., Connell, M. E., & Abduljaleel, N. (2003). Hindi CLIR in thirty days. ACM Transactions on Asian Language Information Processing (TALIP), 2(2), 130-142.

[7] Martynyuk, S. (2003). Statistical Approach to the Debate on Urdu and Hindi (Student Paper).

[8] Rani, R., & Lobiyal, D. K. (2018). Automatic Construction of Generic Stop Words List for Hindi Text. Procedia computer science, 132, 362-370.

[9] Lo, R. T. W., He, B., & Ounis, I. (2005, January). Automatically building a stopword list for an information retrieval system. In Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR) (Vol. 5, pp. 17-24).

[10] Zou, F., Wang, F. L., Deng, X., Han, S., & Wang, L. S. (2006, April). Automatic construction of Chinese stop word list. In Proceedings of the 5th WSEAS international conference on Applied computer science (pp. 1010-1015).

[11] Hao, L., & Hao, L. (2008, December). Automatic identification of stop words in chinese text classification. In 2008 International conference on computer science and software engineering (Vol. 1, pp. 718-722). IEEE.

[12] Yao, Z., & Ze-wen, C. (2011, March). Research on the construction and filter method of stop-word list in text preprocessing. In 2011 Fourth International Conference on Intelligent Computation Technology and Automation (Vol. 1, pp. 217-221). IEEE.

[13] Alajmi, A., Saad, E. M., & Darwish, R. R. (2012). Toward an ARABIC stop-words list generation. International Journal of Computer Applications, 46(8), 8-13.

[14] Kaur, J., & Saini, J. R. (2016). POS Word Class Based Categorization of Gurmukhi Language Stemmed Stop Words. In Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2 (pp. 3-10). Springer, Cham.

[15] Choy, M. (2012). Effective Listings of Function Stop words for Twitter. arXiv preprint arXiv:1205.6396.

[16] Kaur, J., & Saini, J. R. (2016, March). Punjabi Stop Words: A Gurmukhi, Shahmukhi and Roman Scripted Chronicle. In Proceedings of the ACM Symposium on Women in Research 2016 (pp. 32-37). ACM.

[17] Saini, J. R., & Rakholia, R. M. (2016). On continent and script-wise divisions-based statistical measures for stop-words lists of international languages. Procedia Computer Science, 89, 313-319.

[18] Rakholia, R. M., & Saini, J. R. (2016). Lexical classes based stop words categorization for Gujarati language. In 2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Fall) (pp. 1-5). IEEE.

[19] Raulji, J. K., & Saini, J. R. (2017, January). Generating Stopword List for Sanskrit Language. In 2017 IEEE 7th International Advance Computing Conference (IACC) (pp. 799-802). IEEE.

Language Resources

LR#1: Wictionary Top 1900. (June, 2018). In *Wiktionary*. Retrieved July 2, 2019, from https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Hindi_1900.

LR#2: 1000 Most Common Hindi Words. (n.d.). In *1000 Most Common Words.* Retrieved July 2, 2019, from https://1000mostcommonwords.com /1000-most-common-hindi-words.

LR#3: First 100 High Frequency Words in Hindi. (May, 2012). In *Hindi Language Blog*. Retrieved July 2, 2019, from https://blogs.transparent.com/hindi/first-100-high-frequency-words-in-hindi.

LR#4: Top 1000 Words. (n.d.). Retrieved July 3, 2019, from http://home.iitk.ac.in/~prasant/HindiCorpus/word.html.

LR#5: Piotr, Most Common Words by Language (2019). In *GitHub*. Retrieved July 3, 2019, from https://github.com/oprogramador/most-common-words-by-language.

LR#6: Savand, A, Stop Words (2017). In *GitHub*. Retrieved July 5, 2019, from https://github.com/Alir3z4/stop-words.

LR#7: Stopwords ISO, Stop Words-hi (2016). In *GitHub*. Retrieved July 10, 2019, from https://github.com/stopwords-iso/stopwords-hi.

LR#8: Champion, Jason, Extra Stop Words (2016). In *GitHub*. Retrieved July 5, 2019, from https://github.com/Xangis/extra-stopwords.

LR#9: Jha, Vandana, N, Manjunath, Shenoy, P Deepa, & K R, Venugopal (2018), "Hindi Language Stop Words List", Mendeley Data, v1 http://dx.doi.org/10.17632/bsr3frvvjc.1.

LR#10: Hindi Stopwords. (n.d.). Retrieved July 10, 2019, from https://www.ranks.nl/stopwords/hindi.

LR#11: Wikimedia Downloads (n.d.). Retrieved July 11, 2019, from https://dumps.wikimedia.org/backup-index.html.

LR#13: The Open Parallel Corpus (n.d.). Retrieved July 12, 2019, from http://opus.nlpl.eu/.

LR#14: CFILT Hindi Corpus (n.d.). Retrieved July 15, 2019, from https://www.cfilt.iitb.ac.in/Downloads.html.

LR#15: Kunchukuttan, A., Mehta, P. & Bhattacharyya, P. (2018). The IIT Bombay English-Hindi Parallel Corpus. Language Resources and Evaluation Conference.

LR#16: English-Hindi Tourism Text Corpus – EILMT (October, 2016). EILMT Consortia, CDAC Pune. Retrieved July 15, 2019, from http://www.tdil-dc.in.

LR#17: Hindi-English Agriculture & Entertainment Text Corpus ILCI-II (May, 2017). ILCI Consortium, JNU. Retrieved July 15, 2019, from http://www.tdil-dc.in.

LR#18: Hindi Monolingual Text Corpus ILCI-II (June, 2017). ILCI-II, JNU. Retrieved July 15, 2019, from http://www.tdil-dc.in.

LR#19: Hindi-English Health Text Corpus-ILCI (April, 2012). ILCI Consortium, JNU. Retrieved July 15, 2019, from http://www.tdil-dc.in.

An IoT based Home Automation Integrated Approach: Impact on Society in Sustainable Development Perspective

Yasir Mahmood*[1]
Advanced Informatics Department, Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, 54100, Kuala Lumpur, Malaysia
Department of Software Engineering, Faculty of Information Technolgoy, The University of Lahore
Lahore, Pakistan

Nazri Kama[2], Azri Azmi[3], Suraya Ya'acob[4]
Advanced Informatics Department
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia, 54100
Kuala Lumpur, Malaysia

Abstract*—In recent years, due to substantial evolution in the field of consumer electronics, the society is striving to optimize efficiency, energy savings, green technology and environmental sustainability in their daily lives at homes. Most of the people are controlling and monitoring home appliances manually and therefore, facing lots of problems in managing natural resources, cost, effort and security which lead towards an un-comfortable and un-reliable life. Numerous 'intelligent' devices such as smartphones, tablets, air-conditioners, etc. have promoted the key concept of the Internet of Things (IoT) based home automation. Entrenched with technology, these devices can be distantly monitored and controlled over the Internet at home and anywhere in the world. Over the past few decades, global warming has become a severe worldwide challenge. However, sustainable development and green technology play an important role in climate change. The primary purpose of this study is to save natural resources, reduce energy consumption, and to understand the impact of home automation on the society in order to achieve the goal of green technology and environmental sustainability. In this paper, IoT based home automation approach integrated with the smart meter, solar, wind, geothermal renewable energy resources and government green awareness program to extensively optimize the need of energy consumption, security, cost, convenience and cleaner environment for the society is proposed. In addition, a survey was conducted among the target audience for the purpose of identifying and evaluating its least impact on the environment and society in a sustainable development perspective. The results of this survey are statistically analyzed using IBM SPSS statistics version 23. The results revealed that there is a significant impact of home automation on the society thereby contributing to its solution.

Keywords—Internet of Things; smart home; sustainable development; home automation; environment sustainability

I. Introduction

Recent years have witnessed rapid advancements in the Industrial Revolution (4IR) enabling technologies as a chunk of the Internet of Things (IoT) for home automation. The fast-growing thought of Internet of Things (IoT) is used to achieve the goal of highest environmental friendly and green standards of sustainability in order to monitor the home appliances and energy conservation at home environment. The users are gradually offered numerous adapted services like home automation, security, controlling and monitoring, entertainment, healthcare etc. In particular, many home automation research projects have been developed for old and handicapped people to prolong their stay at home for independent living and assisting them in their everyday life [1]. Home automation technology provides security, comfort, power saving, efficiency, go-green technology and environmental sustainability by enabling intercommunication amongst all household appliances through smartphones or other networked devices anywhere in the world. Besides, the home appliances, healthcare and security are the vital features which can be controlled and operated remotely by programs or remote platforms. The usage of smart meters can bring significant impact on home automation systems through which accurate meter readings can be achieved successfully. The smart meter readings can be transferred automatically to the utility companies in order to optimize efficiency. The consumption of home water, electricity, gas, and telecommunications can be monitored and controlled over time. In recent years, smart homes are progressively getting cost-effective and efficient with continued progress in the fast-growing field of IoT technology which links the internet with physical objects globally. Due to the progressive utilization of IoT, a large number of applications have been developed in the domain of transportation, home, building, city automation and healthcare [2]. The IoT has the greatest potential to lower down the environmental load significantly on the society and the users should adopt such technologies which can help them to develop energy efficient buildings and homes. The IoT encourages building constructors and industries in developing more energy efficient buildings and to achieve green home key components like sustainable location, water, energy consumption and indoor environmental quality.

Globally, it is investigated that buildings and homes are consuming approximately 40% annual energy consumption of the total world. This energy is specifically utilized for lighting, heating and cooling systems in building and homes. Some vital steps need to be taken through green awareness strategies of the environmental impact on society using advanced

*Corresponding Author

technologies. Therefore, it is required to investigate reducing the rate of energy consumption and environmental pollution for a cleaner and greener society. One of the ways of reducing energy consumption and environmental sustainability is to automate buildings and homes while achieving good indoor environmental quality life [3].

In this paper cost-effective, a comfortable and reliable environment is created with IoT based home automation technique. Besides, smart meter, solar, wind and geothermal renewable energy resources, and green awareness program are integrated with home automation to minimize demand of energy, optimize the demand response and to achieve the goal of green technology and environmental sustainability for the society. Moreover, a survey is conducted to identify the least impact of home automation approach on the environment and therefore to society.

A. Motivation

Nowadays, society is facing lots of problems in managing their home appliances and energy consumption in terms of affordability, efficiency, effort, reliability and therefore spending uncomfortable and inconvenience life. Furthermore, the chances of robbery and thefts are getting common in the society, hence needs to put serious attention by taking few measures into consideration to ensure the safety of every individual by keeping it as an important element in daily lives. In order to overcome these concerns, the Internet of Things (IoT) plays a significant role for home automation which connects various objects like cellular phones, computers, appliances, for the intent of interaction with each other and an entire world. The main motivation of this paper is to promote green technology and environmental sustainability by developing a smart home using IoT so that Go Green environment can be achieved successfully. We focus our consideration on smart home and study the latest developments in functioning principle of diverse type of wireless communication techniques such as voice recognition, ZigBee, Wi-Fi (IoT based home automation) Bluetooth, and GSM enables energy consumption, green technology and environmental sustainability. Eventually, an integrated home automation conceptual approach is proposed to save natural resources, reduce power consumption and environmental impacts to achieve the goal of green technology and environmental sustainability for the society.

B. Research Contribution

The main contributions of this paper are as below:

RC-1 An integrated home automation conceptual approach is proposed in which smart meter, solar, wind and geothermal renewable energy resources and green awareness program are combined to improve householders' convenience, comfortability, cost, effort, security and to reduce environmental impacts to achieve the goal of green technology and environmental sustainability for the society.

RC-2 A relationship is established between 4IR enabling technologies-Internet of Things (IoT), home automation and renewable energy resources to optimize green technology, energy consumption and environmental sustainability.

RC-3 A comprehensive survey is conducted to validate the impact of an anticipated solution on the society in sustainable development perspectives.

C. Organization

This paper is structured as follows. Section 2 describes a related work to identify gaps in the research area. Section 3 briefly presents a research methodology. An integrated home automation conceptual approach and its architecture described in Section 4. Section 5 explains the impact of the proposed approach on the society. In Section 6 the data analysis and results are discussed. Section 7 presents the conclusion and future research direction.

II. Related Work

The basic idea behind IoT is to make everything around us accessible for monitoring, controlling and communicating over the internet anywhere in the world. However, in order to make a sustainable automation world, all devices should be furnished with sensors, communication tools and techniques so that the users can communicate with the devices globally [4]. Due to substantial energy loads, high costs and limitations on natural resources have upraised the concerns about resource preservation. Researchers are required to investigate buildings and homes power management techniques for energy consumption and conservation. Green IoT emphasizes the smart world with environmental sustainability by reducing energy consumption. A lot of applications [5]-[7] exist related to IoT out of which some are as follows.

Smart Home: With a smart home, the appliances are controlled and monitored by making lifestyle convenient and comfortable. The home devices can remotely communicate and interact with each other.

Industrial Automation: The manufacturing machines are automated therefore helps in finishing their tasks with minimal human involvement. The automated machines are controlled and monitored robotically with an increase in productivity ratio.

Smart Healthcare: The sensors and actuators are embedded in patients for tracking, controlling and monitoring the health conditions of patients, hence enhanced healthcare performance and efficiency.

Smart Grid: Power grid stations are controlled and managed to ensure a continuous supply of energy to the people. Therefore, energy consumption could be enhanced.

Growing issues on global warming, energy consumption and costs have put a great responsibility on researchers for exploring new ways to overcome these concerns. One of the ways is to engage and helping people in climate-friendly behaviours. According to [8], lots of efforts are required to put in on energy consumption on home sectors in particular. For instance, to increase awareness of energy conservation at the workplace and home and to assist people for water, gas and power consumption at different places. Some researchers have proposed home techniques that control and monitor energy consumption and also investigated how residents use home energy tutor system [9]. Few researchers have developed gaming interfaces user-friendly automated application for energy conservation [10].

Information and Communications Technology (ICT) plays a key role in a sustainable environment for the future which is getting one of the most serious challenges of the 21st-century era. The main goal of ICT is to save energy in order to accomplish effective environmental protection for numerous areas including housing, manufacturing, buildings, power consumption, power grid stations, etc. As a result, it is important to highlight that the saving of energy is required where there are wired and wireless network platforms [11]. Therefore, there is a critical need for major research [12] on the reduction of energy use in different types of networks and cellular devices. The green wireless communication technologies [13] help in reducing the impact on the environment.

In recent years, wireless communication technologies are gaining popularity for the use of smartphones which are able to control and monitor the home appliances remotely anywhere in the world. Various home automation systems are using this technology including Bluetooth, Voice Recognition, ZigBee, GSM and Wi-Fi (IoT based).Bluetooth based home automation system is proposed by [14]. The proposed system consists of an Arduino board and a cellular phone which communicate through wireless Bluetooth technology. In this proposed system, the Arduino board is used to connect home appliances through a relay. A Software application is also being developed which permits users for controlling home appliances. The home automation system based on voice recognition is proposed by [15]. The hardware architecture of this system comprises of Arduino UNO and a smartphone. The wireless communication is done via Bluetooth technology. The home appliances can be controlled through the user's voice commands due to built-in voice recognition feature capability of Android smartphone. The working mechanism of this system is that it first converts the voice commands into text messages and then transmits to the Bluetooth module. This module HC-05 is connected with Arduino UNO. ZigBee Based Wireless Home Automation System is proposed and implemented by [16]. This system consists of three modules out of which two modules work as a controller. The microphone is a handheld module, central and appliances are controller modules. Handheld microphone module is based on PC which uses a ZigBee protocol and central controller module. The working mechanism of this system is that it uses Microsoft API for voice recognition and a wireless network is established using RF ZigBee module. GSM Based Home Automation System is proposed by [17] using Global System for Mobile communication (GSM) consists of a GSM modem, smartphone and a microcontroller PIC16F887. This microcontroller is connected with home appliances through relays and a serial communication between GSM modem and a microcontroller PIC16F887 is done with RS232. Moreover, a GSM modem is interfaced with a microcontroller which is used to encode and decode the SMS received for the purpose of execution of the specific commands.

The Internet of Things (IoT) based home automation system is presented by [18] which is used for controlling and monitoring of home appliances based on Internet of Things (IoT) technology. This system consists of controlling devices, smartphone, software application and a micro web server. The system uses three main modules include remote environment, gateway environment and home environment which has the capability to control energy consumption of security, power plugs, door and gate, lightings, air conditioners, refrigerators etc. Another system is proposed by [19] based on the Ethernet system which is used for monitoring real-time tracking and switching information of the devices. The hardware architecture of this system consists of Intel Galileo 2nd generation board, smartphone and Android-based software application to monitor and control the security of homes in case of threats, fire or suspicious activities.

The monitoring and controlling of the home environment using wireless sensors, web server (Apache) and Raspberry Pi is proposed by [20]. The interface of this system is GUI which is easy and convenient for the users to interact with it anytime and anywhere. The hardware architecture of this system contains Arduino microcontroller Atmega 2560 using an Arduino Wi-Fi shield. A relay switch is also being used to send control signals from the Atmega 2560 to the home appliances. The web interface consists of buttons that permit the users to turn ON/OFF a device. The Html, .txt and PHP files are used to store data. The introduction of home automation for controlling and monitoring of home appliances, cooling, security, heating etc. measures over the internet is getting more popular nowadays. One of the beneficial effects it has on the environment is to optimize energy consumption and reduction in energy consumption rates. With the advent of smart metering technology, the consumers are able to monitor and control energy consumption at home which will lead towards cost-effective approach and thus helps the residents to make energy and cost adjustments accordingly [21]. The concept of smart metering integrated with the IoT has given rise to convert buildings and homes into efficient energy-aware environments. The researchers are interested to integrate the IoT in smart grid for the home automation solutions for the intent of saving energy, efficient controlling and to ensure continuous energy supply to the community. The IoT has this potential to provide intelligent, efficient and cost-effective solutions to society for the sake of their convenient, comfortable, better and safe personal lifestyle. The rapid practice of the home automation and with the advent of energy monitors, smart inverters, 21st-century new generation batteries for storage, solar, wind, geothermal energy has joined the IoT for optimizing energy consumption and efficiency. The renewable energy resources such as solar, wind, geothermal etc. are available which plays a significant role in the development of smart home environment [22]. The geothermal energy is currently being used for many smart homes in the world. It utilizes a grounded heat pump for saving energy demand and its consumption over time. Furthermore, smart environmental sensors are extensively being used at home as a key element of the Internet of Things (IoT) which provides monitoring and feedback for all indoor environmental activities. These sensors can measure humidity, temperature, home artificial lighting effects and sunlight.

III. Methodology

This research is engineering and exploratory study with the aim of observing existing solutions and proposed a new approach for home automation for environmental sustainability. In order to identify the research problem, we have used 5W1H (sometimes referred to as Five Ws and How) questions approach whose answers are considered basic in information gathering or problem-solving. In Fig. 1, we can see the problem addressed in this paper by asking 5Ws (who, what, where, when, why) and 1H (how) about a problem taken up for improvement. It has also been observed that non-automated homes have a significant impact on stakeholders, economy and global environment. In order to validate the impact on society, we have used a descriptive approach in nature by designing a survey questionnaire as an instrument to collect required data. The survey is used to validate the impact of home automation on stakeholders, economy and environment in a sustainable development perspective. With the rapid development of IoT, we are primarily focused on improving energy consumption, green technology and environmental sustainability through integrated home automation strategy for the betterment of the society and the global world.

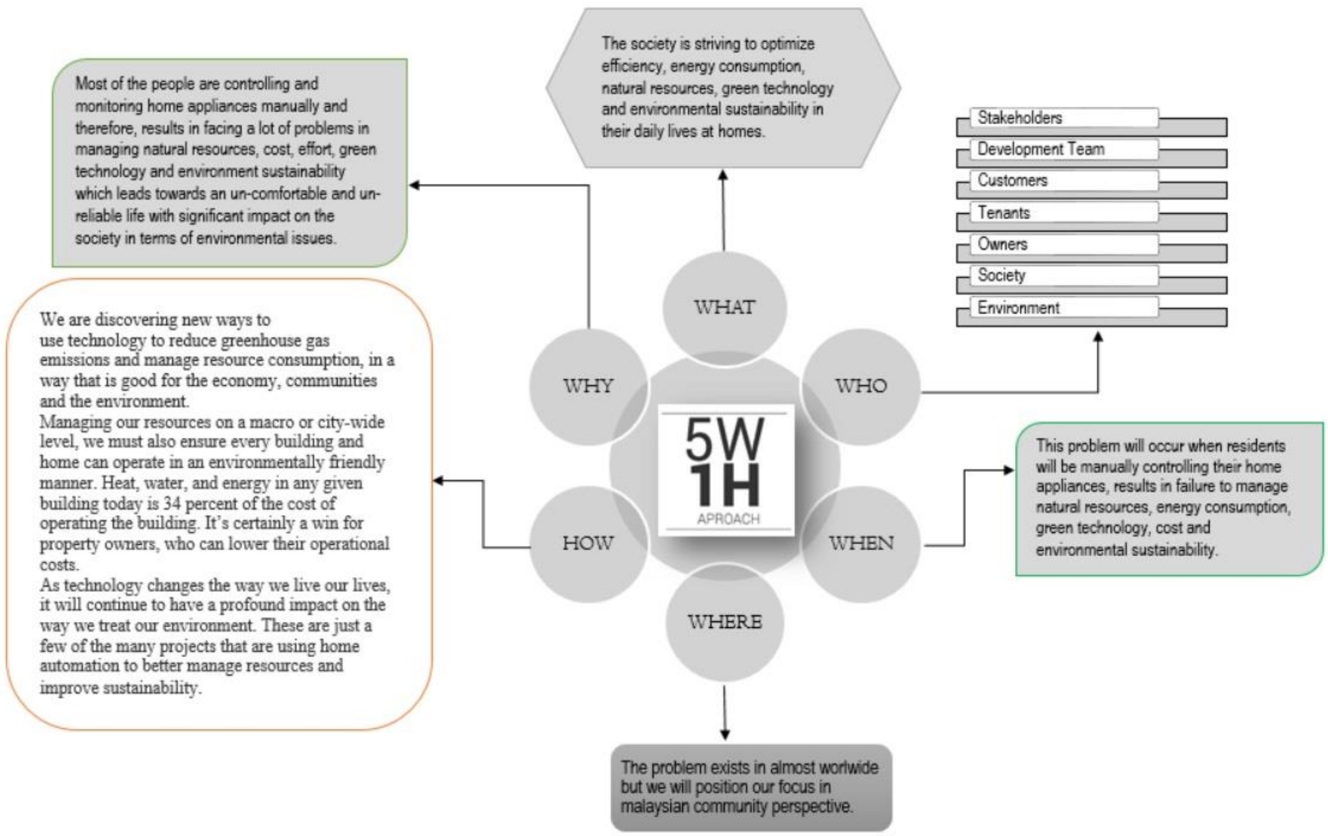


Fig. 1. 5W1H Problem Identification Approach.

IV. Proposed Architecture

The proposed architecture highlights an integrated home automation conceptual approach that optimizes energy consumption considering the environmental sustainable strategies using IoT. The proposed integrated home automation conceptual architecture is shown in Fig. 2 with the aim to optimize effort, cost, security, indoor activities and energy consumption by using renewable resources (solar, wind and geothermal) for the residential sector. Moreover, it is designed and developed by using currently available wireless communication services, equipment and devices through the home network. The residents can control, monitor, and manage energy usage and all home appliances of their choices over the Internet through smartphones using 3G/4G or other network availability options.

The home gateway acts as a data link layer connected with the Internet and home appliances which can be controlled and monitored through the remote client user interface either mobile devices, laptops or other communication devices. The user interface acts as a presentation layer through which the user may interact with. Besides, due to IoT technology, this system will be manageable and accessible from anywhere in the world. The monitoring system is incorporated in this architecture for monitoring the appliances in order to know the energy consumed by such devices. This architecture is a centre of renewable energy systems which are incorporated with home automation technique and smart meters communication for controlling and monitoring home appliances by achieving the aim of optimizing energy consumption and environmental sustainability. The smart meter is used for regular monitoring of energy consumption over time which can be achieved through taking a reading on an hourly or daily basis.

Accordingly, this system provides a significant capability of accessing separate energy system of their own choice by changing the energy system's status for best energy utilization. Several areas are controllable using this approach includes a new generation charging station for the chargeable electrical devices. The geothermal and wind energy produces Alternative Current (AC), a high voltage which needs to be converted into low-voltage for domestic purposes. For economic reasons, the increase and decrease in voltage for transmission and home utilization respectively can only be accomplished by using a step-up and step-down transformer. For this conversion of high-voltage to low-voltage, a step-down transformer is used in domestic consumption. Similarly, in the case of solar energy, it produces direct current (DC) and it needs to be converted in AC which is achieved through using a converter for domestic energy utilization.

In addition, the Government green awareness program is combined to bring awareness among people that could help them by assisting through green awareness campaigns and participatory involvements. This could also be achieved successfully by using social media, advertisements, newspapers, magazines, blogs, interactive simulations, Web portals, etc.

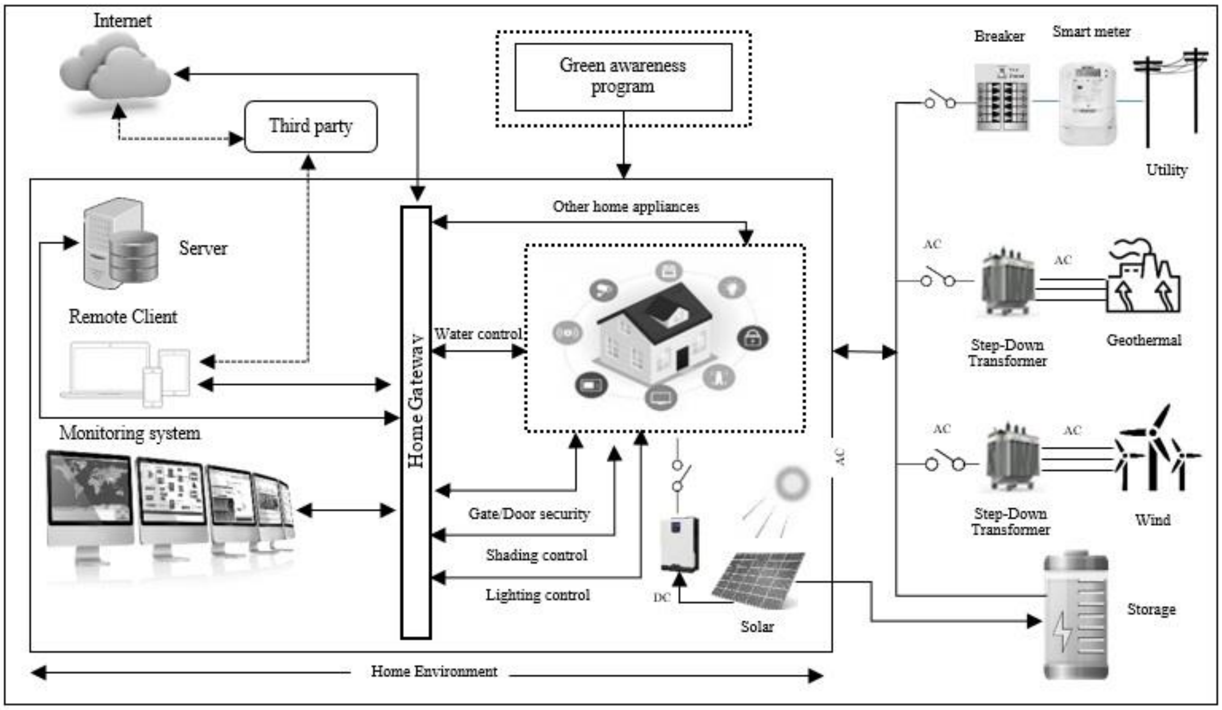


Fig. 2. Integrated Home Automation Conceptual Architecture.

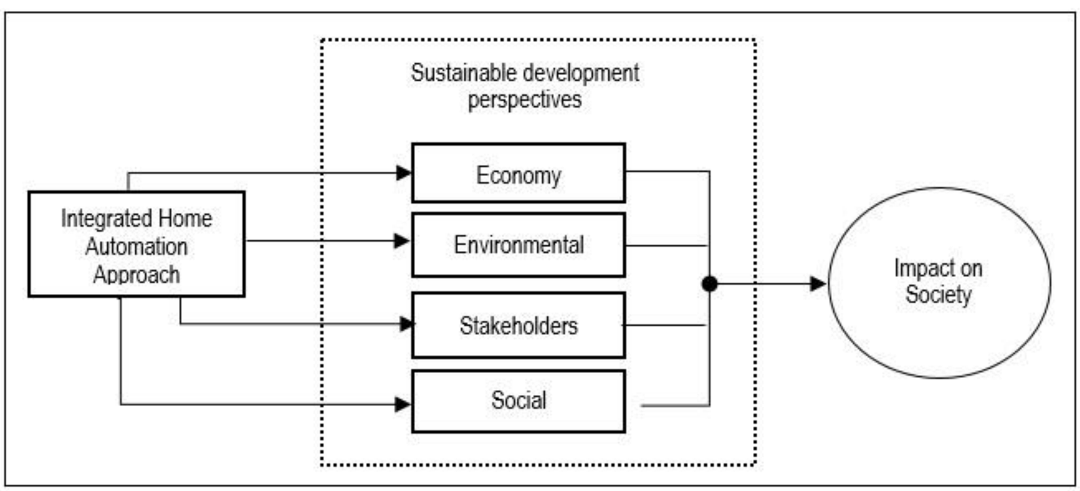


Fig. 3. Structure of Impact Model.

V. Impact of Proposed Approach on Society

A. Structural Impact Model

Fig. 3 illustrates the key structure model. Our research question is: What are the least impacts of home automation on society in a sustainable development perspective? We were interested in knowing the relationship between home automation approach and its impact on the society in view of economy, environment, stakeholders and social. We hoped to show clearly the impact (or lack thereof) of these relationships.

B. Data Collection Methodology

A comprehensive survey was conducted among a target audience that included 20 questions regarding sustainable development perspective through an online questionnaire using Google forms. The measures were tested using a 5-point Likert-type scale ranging from "strongly disagree" to "strongly agree". The purpose of this survey was to validate the least impact of the home automation system on society in a sustainable development perspective. We have divided the population into two main groups. In the first group, we ensured that the participants are computer literate and in the second group we prioritized those who have enough knowledge of environmental engineering and or sciences.

Our work consisted of three phases: 1) creation of the questionnaire, 2) collection of data, and 3) data analysis. In the first phase, a relationship is established between home automation, a green awareness program and environmental sustainability which are the key elements of our proposed solution. The relationship among these attributes indicates the impact of the architecture on the sustainable development and therefore to society.

We then created close-ended questions inquiring on how they can be helpful for establishing a relationship successfully. In the second step, we have used two modes of distributions of the questionnaire 1) manual and 2) online. In the first mode, the questionnaire was distributed manually among the students of these two groups at the Universiti Teknologi Malaysia and outside the community. In the second mode, we sent an online questionnaire link [23] using WhatsApp and E-mails by ensuring that it should be delivered to the respective target audience. In the second phase, the data was collected through Google forms in .csv file format. In the third phase, the .csv file was imported and statistically analyzed using IBM SPSS statistics for the compilation of results. The sample from 112 respondents was checked for missing indicators. It is found that only 3 (2.6%) cases had missing data for one or two of their indicators which appeared randomly. Thus, we decided to keep those cases with mean value substitution. The IBM SPSS statistics version 23 was used to analyze sample responses, standard deviation, skewness and kurtosis for data normality and Cronbach's alpha (α) test for reliability.

VI. Results and Discussion

The total of 112 out of 165 questionnaires was received resulting in 67.87% response rate. The sample was primarily males (71, 63.4%) and females (41, 36.6%) with a mean age of 37.33 (range, 18 to 60 years). A majority of respondents had undergraduate (61, 54.5%) and postgraduate (38, 33.9%) qualification as shown in Table I.

The questionnaire was divided into four categorized by ID's such as ECO (economy), ENV (environmental), SOC (social) and STH (stakeholders) in order to establish a relationship between home automation approach and its impact on the society in sustainable development perspective as shown in Table II. It was observed that for economy (ECO2.1, 2.2 and 2.4), environment (ENV3.1 – 3.4), stakeholders (STH1.1 – 1.8) and social (SOC4.1 and 4.5) perspective, the mean score ranged from 1.94 to 2.75 out of total score of 5 indicating that the respondents were "agreed" on average for impact of the home automation on the society. It was found that for SOC4.2 – 4.4 (highlighted in grey), the mean range showed that the respondents neither agreed nor disagree. Thus, for SOC4.2 we had taken it towards "disagree" that the community should have knowledge regarding the role of home automation for a cleaner environment. It was also observed that for SOC4.3 and SOC4.4 the Row N% is 37.5% and 32.1% respectively which was greater than all other indicators with respect to SOC4.3 and 4.4 hence, showed its significance.

The overall results supported the impact model and revealed that home automation plays a significant role and contribute to sustainable development. The majority of the respondents were agreed that the advancements in home automation technology are an effort to boost the economy. It also provides a greener environment, reduces the usage of energy, effort, natural resources, and saves costs of living and time consumption and therefore contributing to society. Table III reports the data normality of the items used to measure our impact model.

The Standard Deviation (SD) was used to measure the dispersion of the data from the mean values and technically it is volatility. The SD should be used to accurately summarize the descriptive data [24]. The results indicated that the standard deviation for all values was low and it was close to the mean values significantly ranged from ECO2.1 – STH1.8. It has been investigated that the positive and negative skewness and kurtosis values for a perfect normal distribution is between 1.0 and -1.0 and -3 and 3 respectively if the sample size is greater than 100 and less than 200 as a suggested benchmark [25]. In our case, the sample size is 112 thus, the skewness and kurtosis values were between the optimal normal distribution ranges indicated that the results were "disagree" that the community should have knowledge regarding the role of home automation for a cleaner environment. It was also observed that for SOC4.3 and SOC4.4 the Row N% is 37.5% and 32.1% respectively which was greater than all other indicators with respect to SOC4.3 and 4.4 hence, showed its significance. The overall results supported the impact model and revealed that home automation plays a significant role and contribute to sustainable development. The majority of the respondents were agreed that the advancements in home automation technology are an effort to boost the economy. It also provides a greener environment, reduces the usage of energy, effort, natural resources, and saves costs of living and time consumption and therefore contributing to society.

TABLE. I. SAMPLE DEMOGRAPHIC CHARACTERISTICS

		N	N %
Gender	Male	71	63.4%
	Female	41	36.6%
Age	18-40	99	88.4%
	41-60	13	11.6%
	above 60	0	0.0%
Qualification	High School	2	1.8%
	College	6	5.4%
	Undergraduate	61	54.5%
	Post Graduate	38	33.9%
	Doctoral	5	4.5%

TABLE. II. RESPONDENTS' DISTRIBUTION

ID	Strongly agree 1		Agree 2		Neither agree nor disagree 3		Disagree4		Strongly Disagree5		Mean
	Count	Row N %	Count	Row N %	Count	Row N %	Count	Row N %	Count	Row N %	Statistic
ECO2.1	18	16.1%	74	66.1%	14	12.5%	6	5.4%	0	0.0%	2.0714
ECO2.2	8	7.1%	56	50.0%	27	24.1%	16	14.3%	5	4.5%	2.5893
ECO2.4	13	11.6%	57	50.9%	29	25.9%	12	10.7%	1	0.9%	2.3839
ENV3.1	11	9.8%	41	36.6%	27	24.1%	30	26.8%	3	2.7%	2.7589
ENV3.2	13	11.6%	63	56.3%	21	18.8%	13	11.6%	2	1.8%	2.3571
ENV3.3	26	23.2%	67	59.8%	16	14.3%	1	0.9%	2	1.8%	1.9821
ENV3.4	16	14.3%	73	65.2%	20	17.9%	2	1.8%	1	0.9%	2.0982
SOC4.1	15	13.4%	62	55.4%	26	23.2%	9	8.0%	0	0.0%	2.2589
SOC4.2	7	6.3%	29	25.9%	30	26.8%	40	35.7%	6	5.4%	3.0804
SOC4.3	7	6.3%	27	24.1%	33	29.5%	42	37.5%	3	2.7%	3.0625
SOC4.4	7	6.3%	31	27.7%	32	28.6%	36	32.1%	6	5.4%	3.0268
SOC4.5	9	8.0%	40	35.7%	35	31.3%	26	23.2%	2	1.8%	2.7500
STH1.1	28	25.0%	64	57.1%	18	16.1%	2	1.8%	0	0.0%	1.9464
STH1.2	10	8.9%	64	57.1%	26	23.2%	12	10.7%	0	0.0%	2.3571
STH1.3	21	18.8%	64	57.1%	20	17.9%	6	5.4%	1	0.9%	2.1250
STH1.4	19	17.0%	68	60.7%	19	17.0%	6	5.4%	0	0.0%	2.1071
STH1.5	6	5.4%	51	45.5%	29	25.9%	24	21.4%	2	1.8%	2.6875
STH1.6	26	23.2%	61	54.5%	19	17.0%	5	4.5%	1	0.9%	2.0536
STH1.7	24	21.4%	75	67.0%	9	8.0%	3	2.7%	1	0.9%	1.9464
STH1.8	16	14.3%	50	44.6%	32	28.6%	12	10.7%	2	1.8%	2.4107

Table III reports the data normality of the items used to measure our impact model. The Standard Deviation (SD) was used to measure the dispersion of the data from the mean values and technically it is volatility. The SD should be used to accurately summarize the descriptive data [24]. The results indicated that the standard deviation for all values was low and it was close to the mean values significantly ranged from ECO2.1 – STH1.8. It has been investigated that the positive and negative skewness and kurtosis values for a perfect normal distribution is between 1.0 and -1.0 and -3 and 3 respectively if the sample size is greater than 100 and less than 200 as a suggested benchmark [25]. In our case, the sample size is 112 thus, the skewness and kurtosis values were between the optimal normal distribution ranges indicated that the results were significant.

In order to understand and validate the model, whether the constructs in the questionnaire are reliable measures, we examined the reliability and validity of each construct of our survey. For this purpose, a Cronbach's alpha analysis was conducted for the reliability on a total of 20 constructs.

According to Van Zyle et al. [26], α value close to 0 indicates that the results are not reliable and close to 1 suggested that the results are very reliable. As a rule of thumb, α ≥ 0.70 is considered reliable. It was found that the Cronbach's alpha (α) level was 0.796 and Cronbach'alpha (α) based on standardized items was 0.803 as shown in Table IV which indicates that our constructs have a high level of inter-item or internal reliability.

Table V presents the Cronbach's alpha items total statistics which provides with an overall reliability coefficient for a set of variables if any particular item was deleted from the construct. The column "Cronbach's alpha if item deleted" ranged from .779 to .794 (highlighted in grey) indicated "Good" [27] Cronbach's alpha internal consistency for the set of constructs.

The results of the table indicated that the removal of any question would result in a lower Cronbach's alpha means that the set of items are closely related as a group. The removal of any item from the constructs will affect the Cronbach's alpha value leads to the unreliability of the questionnaire constructs.

The bivariate Pearson Correlation was conducted for the validity of variables and to measure the strength and direction of the relationship between variables. It yields a correlation coefficient, r, which measures any linear trend between two or more variables. The benchmark value of 'r' always lies between -1 and 1 [28]. r = 0, indicates no linear relationship, r ¼ 1, means a perfect positive linear relationship and 0 < r < 1 indicates positive linear trend between the variables [29]. Table VI shows a Pearson product-moment correlation for examining the validity and relationship between variables. It was observed that the values at the diagonals are statistically significant at the 0.01 and 0.05 benchmark level hence, indicated its validity.

The linear trend between variables is positively correlated except SOC4.3 (social) and STH1.7 (stakeholder) which is -.11. The relationship between SOC4.1 and 4.2 is .27 means that SOC4.1 (social) was more strongly positively related to SOC4.2 (social). It was found that SOC4.2 explains much more of the variability in SOC4.1. Similarly, the relationship between SOC4.2 and ENV3.3 is .30 indicated significant value means that SOC4.2 (social) was more strongly positively related to ENV3.3 (environment). It was found that ENV3.3 explains much more of the variability in SOC4.2. The share of variability was also calculated. It was found that the coefficient correlation (r) between STH1.3 (stakeholder) and ECO2.1 (economy) is .39 which yields 15.21%. Hence, STH1.3 shares about 15.21% of its variability with ECO2.1.

TABLE. III. MEAN, STD. DEVIATION, SKEWNESS AND KURTOSIS OF KEY MEASURES

ID	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
ECO2.1	112	2.0714	.70665	.834	.228	1.384	.453
ECO2.2	112	2.5893	.97309	.729	.228	.001	.453
ECO2.4	112	2.3839	.86207	.539	.228	.047	.453
ENV3.1	112	2.7589	1.04188	.111	.228	-.938	.453
ENV3.2	112	2.3571	.89902	.818	.228	.425	.453
ENV3.3	112	1.9821	.75911	1.162	.228	3.027	.453
ENV3.4	112	2.0982	.68392	.906	.228	2.723	.453
SOC4.1	112	2.2589	.79123	.501	.228	.035	.453
SOC4.2	112	3.0804	1.04095	-.212	.228	-.809	.453
SOC4.3	112	3.0625	.98896	-.355	.228	-.739	.453
SOC4.4	112	3.0268	1.03506	-.104	.228	-.798	.453
SOC4.5	112	2.7500	.96329	.092	.228	-.742	.453
STH1.1	112	1.9464	.69541	.399	.228	.179	.453
STH1.2	112	2.3571	.79250	.593	.228	-.025	.453
STH1.3	112	2.1250	.80678	.816	.228	1.134	.453
STH1.4	112	2.1071	.73958	.643	.228	.657	.453
STH1.5	112	2.6875	.93028	.393	.228	-.686	.453
STH1.6	112	2.0536	.81472	.817	.228	1.113	.453
STH1.7	112	1.9464	.69541	1.218	.228	3.056	.453
STH1.8	112	2.4107	.92565	.475	.228	-.037	.453
Valid N (listwise)	112						

TABLE. IV. CRONBACH'S A RELIABILITY STATISTICS

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.796	.803	20

TABLE. V. CRONBACH'S A ITEMS TOTAL STATISTICS

Item-Total Statistics					
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
SOC4.1	45.7946	56.435	.364	.301	.787
SOC4.2	44.9732	54.369	.386	.400	.786
ECO2.4	45.6696	55.845	.372	.247	.786
ENV3.3	46.0714	56.571	.371	.399	.787
STH1.1	46.1071	57.826	.291	.409	.791
STH1.8	45.6429	54.304	.456	.390	.781
STH1.6	46.0000	55.514	.429	.472	.783
STH1.7	46.1071	56.871	.384	.428	.786
SOC4.3	44.9911	56.459	.266	.505	.794
SOC4.4	45.0268	55.270	.328	.580	.790
STH1.5	45.3661	55.604	.354	.385	.787
ENV3.1	45.2946	54.786	.357	.493	.788
SOC4.5	45.3036	56.015	.308	.392	.791
ECO2.1	45.9821	57.225	.342	.376	.788
STH1.3	45.9286	54.698	.506	.426	.779
STH1.4	45.9464	55.637	.471	.467	.781
ECO2.2	45.4643	56.521	.268	.314	.793
STH1.2	45.6964	56.664	.343	.326	.788
ENV3.2	45.6964	55.673	.365	.562	.787
ENV3.4	45.9554	57.629	.317	.392	.789

TABLE. VI. CORRELATION MATRIX OF THE CONSTRUCTS

	SOC4.1	SOC4.2	ECO2.4	ENV3.3	STH1.1	STH1.8	STH1.6	STH1.7	SOC4.3	SOC4.4
SOC4.1										
SOC4.2	.27**									
ECO2.4	.10	.23*								
ENV3.3	.08	.30**	.21*							
STH1.1	.29**	-.00	.15	.19*						
STH1.8	.23*	.10	.14	.23*	.23*					
STH1.6	.21*	.05	.26**	.28**	.34**	.54**				
STH1.7	.14	.09	.23*	.36**	.27**	.36**	.45**			
SOC4.3	.08	.25**	.12	-.13	-.02	.04	-.07	-.11		
SOC4.4	.20*	.27**	.19*	-.06	-.07	.09	-.00	.05	.61**	
	STH1.5	ENV3.1	SOC4.5	ECO2.1	STH1.3	STH1.4	ECO2.2	STH1.2	ENV3.2	ENV3.4
STH1.5										
ENV3.1	.40**									
SOC4.5	.26**	.36**								
ECO2.1	-.05	-.04	.12							
STH1.3	.10	.22*	.03	.39**						
STH1.4	.04	.06	.10	.35**	.32**					
ECO2.2	.15	-.06	.01	.09	.24*	.21*				
STH1.2	.04	.09	-.06	.19*	.22*	.35**	.25**			
ENV3.2	.11	-.1	.15	.16	.31**	.39**	.42**	.30**		
ENV3.4	-.02	.05	-.02	.26**	.27**	.26**	.16	.23*	.50**	

**. Correlation is significant at the 0.01 level (2-tailed)

*. Correlation is significant at the 0.05 level (2-tailed)

N=112

VII. Conclusion and Future Direction

The concept of the Internet of Things (IoT) is one of the fundamental technologies to attain the objective of affordable energy at home and environment sustainability. A home automation approach plays a significant role in managing efficiency and to improve the consumption of renewable energy resources. In this paper, an integrated cost-effective, comfortable and reliable approach is proposed with IoT based home automation technique. Besides, smart meter, solar, wind and geothermal renewable energy resources, and green awareness program are incorporated with home automation to minimize the demand of energy and to achieve the goal of green technology and environmental sustainability for the society. This approach also plays a significant role to optimize convenience, comfortability, cost, environment, behaviour and promote awareness among society. The brief reviews on the recent advancements of different type of wireless communication techniques like Bluetooth, Voice Recognition, ZigBee, GSM and Wi-Fi are investigated which may be used during real implementation of the proposed approach. The important element of green awareness program is combined with this approach which could offer assistance to society through a communication medium such as Web portals, wikis, blogs, and interactive simulations. However, the IT industry and society should change their attitude towards positivity for the sake of addressing environmental issues and must take suitable measures to promote and adopt climate-friendly behaviour, strategies and procedures. Moreover, a relationship is established between 4IR enabling technologies-IoT, smart home and renewable energy resources and its impact on the society in a sustainable development perspective. However, a survey is conducted to identify the least impact of the proposed solution on the environment and therefore to the society for its validity.

The survey is statistically analyzed using IBM SPSS statistics in which mean, standard deviation, skewness and kurtosis are used for normal distribution. Furthermore, Cronbach's alpha and Pearson correlation are statistically analyzed for reliability and validity. The results significantly revealed that the usage of home automation has a great impact on society in a sustainable development perspective.

A lot of research still needs to be investigated in this domain. In future work, we recommend that the researchers may explore fog computing and big data for optimizing renewable energy consumption and a large amount of data produced within the IoT.

Acknowledgment

The authors would like to thank and acknowledge each and every individual who has been a source of information, support and encouragement on successful completion of this paper. We are grateful to Universiti Teknologi Malaysia for providing us with all the research facilities required in carrying out this work.

YASIR MAHMOOD
https://orcid.org/0000-0003-1578-5229

References

[1] Z. Zenn Bien. Human-friendly Man-Machine Interaction in Smart Home. Keynote speech at the 3rd International Conference On Smart homes and health Telematic, Sherbrooke, Québec, Canada, July 4-6, 2005.

[2] G. Broll, E. Rukzio, M. Paolucci, M. Wagner, A. Schmidt, H. Hussmann, PERCI: pervasive service interaction with the internet of things, IEEE Internet Computing 13 (6) (2009) 7481.

[3] Abdeen Mustafa Omer, Energy, environment and sustainable development, Renewable and Sustainable Energy Reviews, Volume 12, Issue 9, 2008, Pages 2265-2300, ISSN 1364-0321,

[4] P. Sathyamoorthy, E. C.-H. Ngai, X. Hu, and V. C. M. Leung, ``Energy efficiency as an orchestration service for mobile Internet of Things," in Proc. 7th IEEE Int. Conf. Cloud Comput. Technol. Sci., Nov./Dec. 2015, pp. 1_8.

[5] L. Atzori, A. Iera, and G. Morabito, ``The Internet of Things: A survey," Comput. Netw., vol. 54, no. 15, pp. 2787_2805, Oct. 2010.

[6] C. Perera, C. H. Liu, S. Jayawardena, and M. Chen, `A survey on Internet of Things from industrial market perspective," IEEE Access, vol. 2, pp. 1660_1679, Jan. 2014.

[7] L. Da Xu, W. He, and S. Li, ``Internet of Things in industries: A survey,"IEEE Trans. Ind. Informat., vol. 10, no. 4, pp. 2233_2243, Nov. 2014.

[8] Mankoff, J., Matthews, D., Fussell, S.R. and Johnson, M. Leveraging Social Networks to Motivate Individuals to Reduce their Ecological Footprints HICSS 2007, Hawaii, 2007.

[9] Beckmann, C., Consolvo, S. and LaMarca, A. Some Assembly Required: Supporting End-User Sensor Installation In Domestic Ubiquitous Computing Environments Ubicomp 2004, Springer-Verlag, 2004.

[10] Arroyo, E., Bonanni, L. and Selker, T. Waterbot: exploring feedback and persuasive techniques at the sink CHI 2005, ACM, Portland, Oregon, USA, 2005.

[11] O. Orhan, D. Gunduz, and E. Erkip, "Energy harvesting broadband communication systems with processing energy cost," Wireless Communications, IEEE Transactions on, vol. 13, no. 11, pp. 6095–6107, Nov 2014.

[12] K. Davaslioglu and E. Ayanoglu, "Quantifying potential energy efficiency gain in green cellular wireless networks," Communications Surveys Tutorials, IEEE, vol. 16, no. 4, pp. 2065–2091, Fourth-quarter 2014.

[13] B. Wang, Y. Wu, F. Han, Y.-H. Yang, and K. Liu, "Green wireless communications: A time-reversal paradigm," Selected Areas in Communications, IEEE Journal on, vol. 29, no. 8, pp. 1698–1710, September 2011.

[14] R. Piyare and M. Tazil, "Bluetooth based home automation system using cell phone," Consumer Electronics (ISCE), 2011 IEEE 15th International Symposium on, Singapore, 2011, pp. 192-195.

[15] S. Sen, S. Chakrabarty, R. Toshniwal, A. Bhaumik, "Design of an intelligent voice-controlled home automation system", International Journal of Computer Applications, vol. 121, no.15, pp. 39-42, 2015.

[16] H. AlShu'eili, G. S. Gupta and S. Mukhopadhyay, "Voice recognition based wireless home automation system," Mechatronics (ICOM), 20114th International Conference On, Kuala Lumpur, 2011, pp. 1-6.

[17] R. Teymourzadeh, Salah Addin Ahmed, Kok Wai Chan and Mok Vee Hoong, "Smart GSM based Home Automation System," Systems,Process & Control (ICSPC), 2013 IEEE Conference on, Kuala Lumpur, 2013, pp. 306-309.

[18] R. Piyare, "Internet of things: ubiquitous home control and monitoring system using android based smartphone", International Journal of Internet of Things, vol. 2, no. 1, pp. 5-11, 2013.

[19] Gupta, P., & Chhabra, J. (2016). IoT based Smart Home design using power and security management. 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH). doi:10.1109/iciccs.2016.7542317.

[20] D'souza, M., Wilfred, N., Pereira, R., Rayen, T., & Telgote, A. (2017). Home automation using Internet of Things. 2017 International

Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). doi:10.1109/icecds.2017.8389498.

[21] L. Song, Y. Xiao, and M. van der Schaar, "Demand-side management in smart grids using a repeated game framework," Selected Areas in Communications, IEEE Journal on, vol. 32, no. 7, pp. 1412–1424, July 2014.

[22] M.W. Ahmad, M. Mourshed, D. Mundow, M. Sisinni, Y. Rezgui, Building energy metering and environmental monitoring – a state-of-the-art review and directions for future research, Energy Build. 120 (2016) 85–102.

[23] https://tinyurl.com/y67bfyg9.

[24] Barde MP, Barde PJ. What to use to express the variability of data: Standard deviation or standard error of mean?.Perspect Clin Res. 2012;3(3):113-6.

[25] 2005 American Statistical Association Journal of Business & Economic Statistics January 2005, Vol. 23, No. 1 DOI 10.1198/073500104000000271.

[26] S.M. van Zyle, N. Heinz, D.G. Nel, On the distribution of the maximum likelihood estimator of Cronbach's alpha, Psychometrika 65 (2000) 271–280.

[27] DeVellis, R.F. (2012). Scale development: Theory and applications. Los Angeles: Sage. pp. 109–110.

[28] Puth, M.-T., et al. (2014). "Effective use of Pearson's product–moment correlation coefficient." Animal Behaviour 93: 183-189.

[29] Schober, P., et al. (2018). "Correlation Coefficients: Appropriate Use and Interpretation." Anesth Analg 126(5): 1763-1768.

APPENDIX

The following twenty sustainable development attributes were measured based on a five-point Likert scale. The scale was weighed as strongly agree, agree, neither agree nor disagree, disagree and strongly disagree.

No.	Constructs
SOC4.1	I have sufficient knowledge about the current environmental issues.
SOC4.2	I believe that the community has enough knowledge regarding the role of home automation for a cleaner environment.
ECO2.4	Rapid advancements in home automation technology is an effort to boost the economy can contribute to environmental degradation.
ENV3.3	Personal awareness of home automation is an important factor in the protection of the environment.
STH1.1	Home Automation System (HAS) through Internet of Things (IoT) contribute to a better quality of life
STH1.8	Do you think it gives more freedom of mobility to society?
STH1.6	According to your opinion, it gives people more control over and productive in their daily lives.
STH1.7	Home automation with renewable energy is beneficial for energy providers in order to maintain electricity and gas grid stations.
SOC4.3	Home automation through IoT is not helpful because they do not explain things in terms I understand.
SOC4.4	The home automation system is not designed for use by ordinary people.
STH1.5	People are too dependent on home automation through IoT technology to do things for them.
ENV3.1	Too much home automation through IoT technology distracts people to a point that is harmful.
SOC4.5	I do not feel confident to implement automation with a place that can only be reached physically.
ECO2.1	It reduces the usage of energy, effort, wastage of natural resources, saves costs of living and convenience, and time consumption.
STH1.3	Do you think it helps to achieve the goal of a sustainable environment?
STH1.4	According to your understanding, it is the way for a cleaner and greener environment for future generations.
ECO2.2	Do you believe it provides low maintenance cost on home appliances, hence spent income on other services?
STH1.2	It helps in engaging and focusing on social activities while reducing home tasks.
ENV3.2	Do you think it saves natural resources and minimizes global warming?
ENV3.4	It brings significant impact on less energy consumption and cleaner environment through solar, wind and geothermal renewable energy.

Crime Mapping Model based on Cloud and Spatial Data: A Case Study of Zambia Police Service

Jonathan Phiri[1], Charles S. Lubobya[3]
Department of Electrical and Electronic Engineering
University of Zambia
Lusaka, Zambia

Jackson Phiri[2]
Department of Computer Science
University of Zambia
Lusaka, Zambia

***Abstract*—Crime mapping is a strategy used to detect and prevent crime in the police service. The technique involves the use of geographical maps to help crime analysts identify and profile crimes committed in different residential areas, as well as determining best methods of responding. The development of geographic information system (GIS) technologies and spatial analysis applications coupled with cloud computing have significantly improved the ability of crime analysts to perform this crime mapping function. The aim of this research is to automate the processes involved in crime mapping using spatial data. A baseline study was conducted to identify the challenges in the current crime mapping system used by the Zambia Police Service. The results show that 85.2% of the stations conduct crime mapping using physical geographical maps and pins placed on the map while 14.8% indicated that they don't use any form of crime mapping technique. In addition, the study revealed that all stations that participated in the study collect and process the crime reports and statistics manually and keep the results in books and papers. The results of the baseline study were used to develop the business processes and a crime mapping model, this was implemented successfully. The proposed model includes a spatial data visualization of crime data based on Google map. The proposed model is based on the Cloud Architecture, Android Mobile Application, Web Application, Google Map API and Java programming language. A prototype was successfully developed and the test results of the proposed system show improved data visualization and reporting of crime data with reduced dependency on manual transactions. It also proved to be more effective than the current system.**

Keywords—Zambia police; web application; mobile application; cloud model; crime mapping; spatial data

I. INTRODUCTION

Challenges in preventing and reducing crimes are what most governments around the world are struggling to deal with, every family and business have been directly affected by different kinds of crimes like robberies, vandalism, burglaries, sexual and other crimes [1].Crimes affect the quality of life, economic growth, and reputation of a nation. There is need for the law enforcements to take tough preventive measures to reduce crimes in communities [2]. In Zambia the Zambia Police is considered as the main law enforcement agency mandated to enforce law on Zambian citizens and combat crime thereby playing a critical role in the Zambian criminal justice system [3]. A criminal justice system comprises of government institutions mandated to detect and mitigate crime, it focuses on how criminal cases flow from the time they are reported and investigated up to when they are disposed off. A well established and effective criminal justice system is the key to the reduction of crime in a nation [4] . Ordinary citizens and communities in a Government expect the criminal justice system and its general capacity to not only protect the communities and deal with criminal offenders but also interact with different various parties including victims, witnesses, accused as well as criminal justice professionals [5], therefore the Zambia police being the first to have contact with these people can be regarded as the gate keeper of the Zambian justice system. The Zambia Police was established in 1891 under British South African Company known as Northern Rhodesia police force, and later in 1964 upon attainment of independence was established then under Article 103 (3) of the constitution and now under Article 193 (2) of the 2016 amended constitution of Zambia and also under the Zambia Police amendment act number 30 of 2016 of the laws of Zambia, the name was changed from Northern Rhodesia to Zambia Police force which later in 1994 changed to Zambia police service. Article 193 (2) of the 2016 amended constitution clearly outlines the roles and functions of the Zambia police service, it mandates the agency to ensure protection of life and property, preservation of peace, maintenance of law and order, upholding bill of rights and most importantly detect and prevent crime [6].One of the key strategies used to detect and prevent crime is crime mapping. The technique involves the use of geographical maps to help crime analysts identify and profile crimes committed in different residential areas, as well as crafting best methods of responding [7]. It facilitates visual and statistical analysis of spatial crime data for a specific area by linking it with geographical variables like bars, schools, streets and others. Crime does not spread across the space evenly or equally but rather clumps on some specific areas while absent in other areas. Crime mapping is devoted at identifying high crime areas or neighborhoods also known as hotspots, hotspots are areas with high criminal activities [8]. Proactive policing pushes police officers to identify areas with high concentration of crimes, determine what causes these concentrations and find methods of reducing these concentrations [9]. A well and clearly visualized crime hotspot map significantly helps police officers in aiding threat visualization, police resource allocation and crime prediction. The development of geographic information system (GIS) technologies and spatial analysis applications coupled with cloud computing have significantly improved the ability of crime analysts to perform this crime mapping function [10]. However, the Zambia Police Service is still using the manual and traditional way of mapping crimes, crime data management and crime data visualization. All crime data and records are stored in books and papers. Crimes are mapped using physical geographical maps and pins to establish the location of crime areas. It is so complex to keep track of criminal activities and statistics in crime hotspot areas using paper based system. The manual and paper based crime mapping system that is in place does not provide the needed efficiency and effectiveness to the management of crime maps and crime data.

The focus of the study is on computerizing the crime mapping processes by using cloud architecture and spatial data.

This paper is organized as follows: The second section is literature review which covers technologies used in crime mapping and cloud computing, it also covers the related works which looks at the systems that have previously been designed and implemented to solve challenges of the similar nature. Third section presents the methodology, fourth section brings out the results, the fifth section presents the discussion of the baseline study and development of the prototype crime mapping system, the sixth section presents the conclusion and the last section presents recommendations and future works.

II. Literature Review

A crime is defined as the breach of criminal law that govern a particular geographical area, the criminal law that aims at protecting the lives, property and rights of citizens within a particular jurisdiction [11].In addition ,Harries [12] describes a crime in four dimensions; Legal – a law must be broken, Victim – someone or something has to be targeted, Offender – someone has to do the crime and Spatial – a crime has to happen at a place somewhere in space and time, commonly referred to as crime spatial data. Spatial data is data that describes the location, shape and relationship of geographic features [13].In criminology it is called spatial crime data because it contains geographical referenced attributes like geographical coordinates (longitude and Latitude) that can be used to establish the exact location of an object or crime incident on the map [14].Crime mapping seeks to answer the question of “where”?" for example, “where does crime happen?” or “Where should we focus efforts to catch a serial killer?” or “Where should we build a new police station to fight the crime?” or “Where is crime highest?” by analyzing the crime spatial data provided from various sources [15].Daglar and Argum [16]in their paper highlighted that a place of a crime and any other geographic information connected with a criminal incident can give a lot of information about characteristics of possible criminals, it can also assist in designing of assessment and prevention programs for the related crime incidents. They added that deciding a policing method is always influenced by some facts of place such as jurisdictions, zones, and incident locations are all related to geography. This section will present the history of crime mapping, crime mapping technologies, cloud computing and related works.

A. History of Crime Mapping

The use of traditional crime pin maps for data visualization dates back in early 1830s.The crime map was a representation of pins stuck on it, the pins on the map were useful for showing where crimes occurred [12].There are three schools that were recognized and pointed out in the study by Philips [14],the first was called cartographic/geographic school which dominated between 1830s and 1880s originated from France and later spread to England. In this work, government begun to collect social data, the intended purpose was to center on the influence of variables such as wealth and population density on levels of crime. The second was called Typological school which dominated between 1880s and 1900s, it focused on the relationship between the mental and physical characteristics of people and crime. The third was called Social ecological school which has dominated from 1900s up to date, it concentrated on the geographical variations on social conditions assuming that they were related to crime patterns. It focused primarily on recognizing and classifying areas in the cities with similar social characteristics [12].

B. Crime Mapping Technologies

The manual and traditional pin maps had a number of limitations among them were loss of data as they were updated and also the maps were static, they could not be manipulated or queried. Mapping of crimes has become a primary function in law enforcement agencies, the advancements in computing have facilitated the development of geographical system and spatial technologies. There are three main categories of spatial technologies that can be used for crime mapping [17];

1) Open source maps: These are maps that available over the internet for free, examples of common open source maps are google maps, Bing maps and open-street maps. Crime analysts are able to perform basic crime mapping duties at free cost. The disadvantage of open source maps is that crime analysts cannot perform complex queries.

2) GIS-Software: These tools are commercial, they provide the necessary geospatial analytical functions needed for basic, intermediate and advanced mapping queries. The main disadvantage of GIS-software is the high cost of acquiring a license, it also requires special training.

3) Online dashboard: In this category, law enforcement agencies outsource the crime mapping duties from agencies outside the law enforcement. Agencies like Bair Analytics can provide crime mapping services at a nominal fee. The main advantage of outsourcing is that it reduces the workload of geocoding and data organization. It also provides better data storage and security as data is stored via cloud–storage.

The advancement in computing has also led to the use of artificial intelligence technologies like machine and deep learning models to predict future crime occurrences [18] , this approach is known as predicting policing. Predictive policing involves the use of analytical techniques to identify either likely places of future crime scenes or past crime perpetrators, by applying statistical predictions [19].

Machine Learning can be described as a field of computer science that evolved from studying pattern recognition and computational learning theory in artificial intelligence. It is the learning and building of algorithms that can learn from and make predictions on data sets [20]. It is regarded as the field of study that gives computers the ability to learn without being explicitly programmed. The main purpose of machine learning is to learn from the data without any human intervention [21]. There are three main categories of machine learning algorithms namely supervised, unsupervised and semi-supervised machine leaning algorithms. The algorithms in the Supervised Machine learning are those that need external assistance, in this category, the provided input dataset is divided into train and test dataset. The train dataset has output variable which needs to be predicted or classified. All algorithms learn some kind of patterns from the training dataset and apply them to the test dataset for prediction or classification [22]. There are three types of supervised machine learning algorithm, the first is Decision trees - Decision tree is used mainly for classification purpose where attributes of groups are sorted based on their values. Each decision tree consists of nodes and branches, a node represents attributes in a group that is to be classified and a branch represents a value that the node can take. The second is Naïve Bayes -This algorithm manly focuses on classification of text, it mainly depends on conditional probability of events. The third is Support Vector Machine (SVM)-It mainly focuses on calculation and classification of margins, it basically draws the margins between the classes. The algorithms in the Unsupervised Machine Learning learn few features from the input pattern or data, the learning data is divided into different clusters hence referred to as a clustering algorithm. When new data is introduced, it uses the previously learned features to recognize the class of the data. It is mainly used for clustering and feature reduction [23]. There are two types of unsupervised machine learning algorithms, the first one is K-Means Clustering- is a type of unsupervised learning technique that automatically create groups when initiated. The items which possess similar characteristics are put in the same cluster creating K-district clusters hence called K-means clustering algorithm. The second is Principal Component Analysis or PCA, in this algorithm, the dimension of the data is reduced to make the computations faster and easier. The algorithm in the semi-supervised machine is the combination of both the power of supervised and unsupervised learning. It is mostly ideal and fruitful in areas such as data mining where there is a presence of unlabeled data and getting the labeled data is deemed a tedious process. Some of the semi-supervised models include generative model, self-training model and transductive model.

Deep Learning is also utilized to predict crimes using different techniques [24]. It is regarded as the newly introduced area of machine learning and artificial intelligence comprising of multiple hidden layers of artificial neural networks. Its roots are derived from an Artificial Neuron Network (ANN) introduced by Kunihiko Fukushima in 1980. An ANN can be described as an interconnected network of processing units emulating the network of neurons in the brain [25]. Deep learning has two main categories namely deep neural networks and convolutional neural networks. A Deep Neural Network (DNN) is an artificial neural network (ANN) that has multiple hidden layers of units between the input and output layers capable of modeling complex non-linear relationships [26]. Convolutional neural networks (CNNs) are an artificial neural network that use convolution in place of general matrix multiplication in at least one of their layers. They use used tied weights and pooling layers, this allows them to take advantage of the 2D structure of input data. They can be used in both image and speech applications [27].

C. Cloud Computing

The other technology that has emerged, developed so fast and contributed to the spatial data technology is cloud computing. The evolution of GIS technology and of spatial information acquisition technology have led to more and more collection of spatial data through various approaches for different services like emergence services, crime mapping and other reasons, this has caused the demand for high information processes and computing environment [28]. Cloud computing has emerged as a technology primarily focusing on large scale resource sharing and low cost for big data storage technology. Cloud computing has proved to be an emerging technology capable of providing business models for organizations to utilize deferent computing services at a minimum cost. Cloud is defined as a metaphor describing a web as a space where computing has been preinstalled and exist as a pool of services such as information, infrastructure, applications, storage and processing power on the web ready to be shared [29]. The cloud computing architecture consists of the front end and back end components [30] as shown in Fig. 1. An internet connectivity is required to effect communications between the client and the backend.

The front end is referred to as a client part of cloud computing system which consists of interfaces, networks and applications that are required to access the cloud computing platforms. The back End also referred to as a cloud itself, consists of all the resources required to provide cloud computing services. It comprises of huge data storage, virtual machines, security mechanism, services, deployment models, servers and many others that can be accessed securely accessed over the internet or private connection [32]. Cloud computing offers three key service models namely Software as a Service (SaaS), Platform as Service (PaaS) and Infrastructure as a Service (IaaS) [33] as shown in Fig. 2.

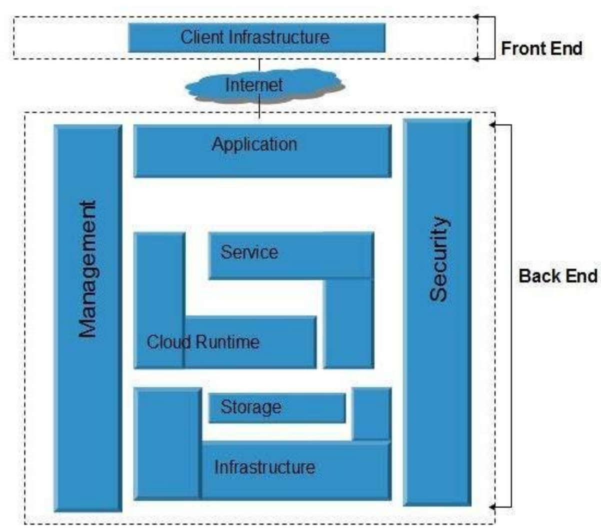


Fig. 1. Cloud Computing Architecture [31].

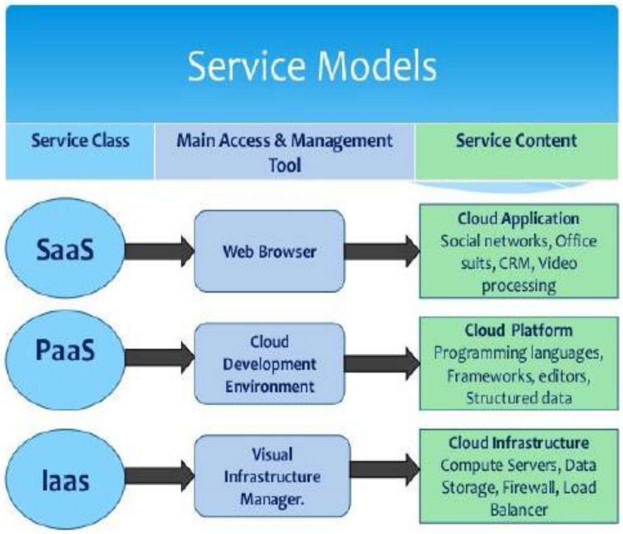


Fig. 2. Cloud Service Model [34].

Software as a service (Saas) delivery model in which software applications are hosted by a vendor or service provider and made available to customers over a network, typically the Internet, the clients can purchase and install the application onto personal computers. Platform as a service (Paas) is a delivery model in which clients are provided with the environment for development and deployment of web based applications using required tools like already created library, pattern, services, programming language and many others. Clients cannot manage the servers, operating system, storage and network but they have control over the deployed web applications including re-configuring the applications. Infrastructure as a service (IAAS) delivery model is where users are allocated with computing resources like operating system, servicers, networks, storage media and others in order to run their applications [34]. There are five models in which cloud computing can be deployed namely public cloud, private cloud, community cloud and hybrid cloud models [35]. The public cloud is where all the computer systems and services are accessible and available to the general public or a large industry group. The private cloud is where computer systems and services are available and restricted to only an organization and that organization is given greater control, security and privacy. The hybrid cloud is where both public and private cloud are combined to offer a common service, on-critical activities are performed using public cloud and critical activities are performed using private cloud. It is mostly used for archiving and backing up of data by replicating local data to a public cloud. The community cloud is used and controlled by a group of organization from specific communities that have similar and common computing requirements and interests [35].

With the continuous historic advancement in technology and high demand for data sharing, cloud computing has been recognized as the most flexible delivery model for ICT resources [36]. According to [37]in their paper they proposed an approach which employs cloud-based service to solve the big spatial data technology in emergence management for better spatial analysis. Mwansa and Phiri in their paper [38] developed a model for the inventory system based on quick response and cloud computing integrated with mobile application for real-time capture of grain bags brought in by farmers at the setline depot. The results of the system showed an improved and acceleration of grain stock statistics in real-time.

D. Related Works

Most Police stations and other law enforcement agencies in developed countries like USA and UK already migrated from traditional pin maps to computerized crime mapping systems [39]. The police in most developing countries like Africa are not equipped with infrastructure like GPS and GIS technologies for mapping of crime. The infrastructure is not only expensive to acquire but also time consuming when setting it up, it requires trained people to operate and manage it. But the benefits of viewing data in the form of a map are massive as compared to grasping tabular or manual data of crime incidents [40]. In Africa, South Africa is considered as more progressive than any other country in Africa in terms of usage of ICTs. In 2000 the South African Police Service (SAPS) through the crime information center developed the National crime GIS database that could allow SAPS to link crime statistics with police boundaries as well as exploring the relationship between crime types and social demographic variables through the use of multivariate statistical techniques. In addition, the author highlights that currently the predominant information system used by SAPS is Case Administration System (CAS) which is regarded as the primary source of information on victims and offenders. It is integrated with a case docket management system that gathers information at a police station level on crime cases such as address and time of crime. The challenge is that CAS is not yet fully accessible country wide and it has not been linked with any GIS technology making crime spatial description and interpretation of criminal activities impossible [41].

Tong in her paper [42] also proposed a crowdsourcing based crime mapping system, it’s a system based on cloud architecture integrated with iPhone mobile application. The system composed of a server running in the cloud and a client application that includes a website and iPhone mobile application to enable users to interact with crime contents. The system gathered both crime reports from the crowd together with crime contents and displayed them on a crime map. Users of the system are able to review and comment on crime incident on a digital crime map.

Singh et al. [43] in their paper proposed and developed a cloud GIS crime mapping blog that could be used by the police for crime mapping. The system generates daily, weekly and monthly crime maps which would help to identify crime patterns and clusters. The crimes in the blog are categorized into theft, murder, snatching and other classes, the system was developed on a cloud architecture using Google cloud and Google maps as source of spatial data. The system uses RSS (Really Simple Syndication) feeds from various news websites as a source of crime data, meaning that crime related news is manually extracted from the collection of news and formatted into desired format with attributes like location, type of crime, details and link to the news web page. The location data is further transformed into georeferenced data, the process commonly known as geocoding. Geocoding is described as the process of converting or transforming a description of a location such as pair of coordinates, an address into a location on the earth's surface and used to point a location in the GIS digital map [44]. Depending on the details of the address, this could be a specific building, the center point of a road, or the center point of an area [45]. Geocoding technique requires special skills and training. The limitation of this system is that the location of the crime specified in the news is generalized therefore the generated point location is not very accurate.

Mwiya et al. [46] proposed and developed a public crime reporting system for the Zambia Police, a cloud & GSM based android mobile application (prototype) to help the general public to only report crime cases using mobile devices. The limitation of this system is that it does not map crime areas.

Some of the solutions provided by applications in the related works would be of great benefit if adopted in the crime mapping model for the Zambia Police. Due to unavailability of digital historical crime dataset from the Zambia Police, the source of crime data for the proposed system is the live crime reports from the general public, to achieve this, a crime reporting component would be added to the system consisting of a web application and mobile application. In order to enhance accuracy, users would specify crime location by selecting a name of the location from the Google map, then the system saves the georeferenced data into the cloud database. The benefit of this approach is that it does not require crime data to be geocoded to produce crime maps as everything would be done by the system.

III. Methodology

This section of the study presents the materials and methods used to conduct the baseline study, this includes methodology used to design the model then followed by mapping of business process.

A. Baseline Study

The purpose of the baseline study was to identify challenges in the current crime mapping system used by the Zambia Police. To archive this, different materials & methods were used;

1) Data collection: Primary data was collected from police officers based in Lusaka working in the CID and VSU departments through the use of interviews & structured questionnaires. Further, general members of public were also interviewed through self-administered questionnaires. Open-ended and closed-ended questions were included in the questionnaires to capture both the qualitative and quantitative responses.

2) Population and sample: The population considered for this study was the Lusaka based police officers. Purposive sampling method was used to select 88 police officers from ten different police stations across Lusaka District. Further, a combination of random and convenient sampling method was used to select members of public who were visiting the police stations for various reasons.

3) Data analysis and presentation: Data was entered, analyzed and presented using the Statistical Package for Social Scientists (SPSS) version 16. The analysis was mainly descriptive in nature.

B. Model Design Methodology

The proposed model was designed using Cloud Architecture, Android Mobile Application, Web Application, Google Map API and java Programming Language. The web application consists of a client and server architecture where a client is a browser and server is the cloud firebase thus include firebase-database for data storage.

C. Mapping Business Processes

Fig. 3 shows the crime mapping business processes for the current crime mapping system,

As shown in Fig. 3, crime mapping process begins at the time when a crime is officially reported by a member of public. The case including all the details of the complainant and suspect if any are recorded into an occurrence book by an officer on duty at the front desk known as inquiries. Thereafter the case is forwarded to the CIO for review, approval and assigning of case to an investigator. The investigating officer investigates & analyze a crime by placing a pin on a geographical map depicting its location. The investigating officer forward the case details to the Records/Statistics officer who enters the case into the crime register for record purposes. The records/Statistics officer generates crime statistics and share the information with an investigation officer and others.

The Proposed automated crime mapping business processes are derived from the current business processes presented in Fig. 3. The proposed model is designed in two parts, the mobile application to be used by the general public to report crimes and also the Web application to be used by the police to not only capture and view crime reports but also generate crime statistics and crime maps. Fig. 4 shows the proposed web application.

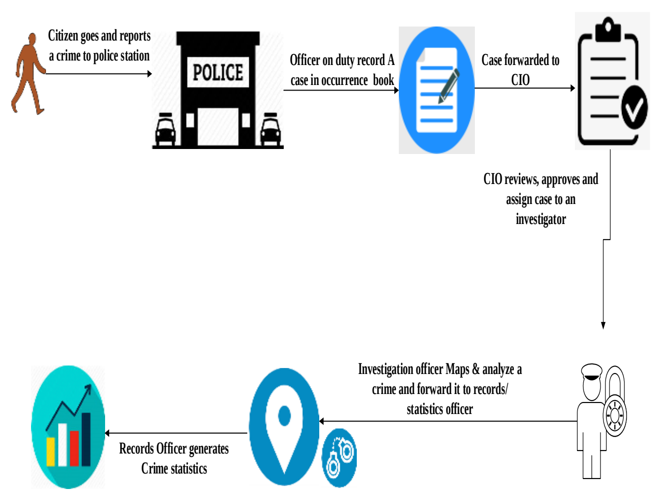


Fig. 3. Current Crime Mapping Business Processes.

As shown in Fig. 4, a crime is directly reported by a member of the public, the officer on duty records the crime case into the system, the GPS satellite through Google maps captures not only the actual location of crime but also residential address of both the complainant and suspect, the case details including geo-referenced data will be saved into the cloud database. The case is automatically forwarded to the CIO for assigning, upon assigning the case to the investigator, the case is added to the crime register and automatically updates the crime map. Both the crime investigator & statistics/records officers will be able to view and generate crime maps & reports. Fig. 5 shows the proposed mobile application business processes.

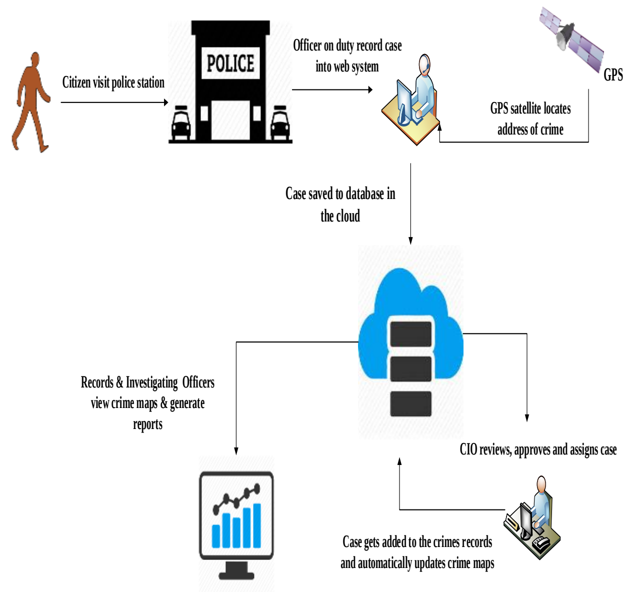


Fig. 4. Proposed Business Processes - Web Application.

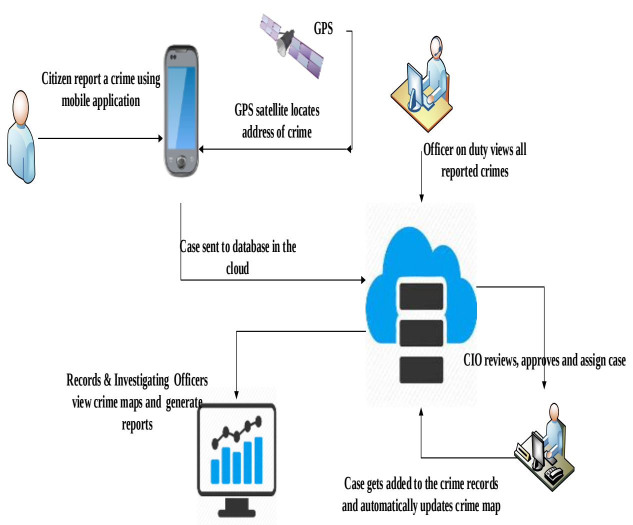


Fig. 5. Proposed Business Processes – Mobile Application.

The mobile application allows the citizen/user to report crime case using a mobile device. The police will use the web platform to view the reported crimes and map them. The citizens are also able to view the status of the case they reported.

D. *Proposed System Architecture*

The proposed system architecture shown in Fig. 6 utilizes the private cloud infrastructure where ZAMTEL the largest telecommunications company in Zambia provides the MPLS network while Zambia Police provides the private cloud services. The MPLS backbone comprises of fiber and microwave.

As shown in Fig. 6, Zambia Police exclusively operates computing resources in the cloud in which different servers are installed such as Email server, Database server, Application server, Web server, Real-time communication server and many others. The rationale herein is to make sure the system is accessed by different police stations located across the country thus include the general public who are the mobile application users with less convenience of configuring hardware, software and security of information. The proposed architecture will allow Zambia Police to have complete control of the system including how data is managed and what security measures are to be put in place.

E. *System Modelling*

1) Use case: The use case depicts how the users or actors interact with the system, in the proposed system there is a web and mobile application use cases. The use case for web application is shown in Fig. 7.

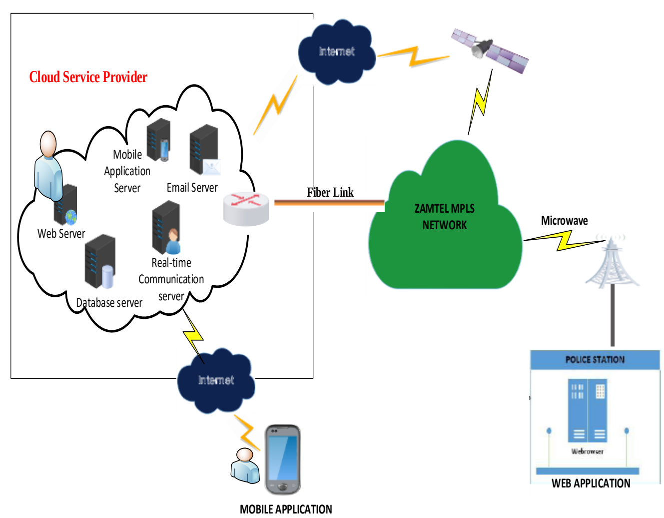


Fig. 6. Proposed System Architecture.

As shown in Fig. 5, the main actors in the web system are; officer on duty, CIO, investigating officer & records/statistics officer. The users at the web application first log into the system and then perform transactions like creating and updating crime cases, view reported crimes and also view generated crime maps.

The use case for mobile application is shown in Fig. 8.

As shown in Fig. 6 the main actors in the mobile app are the citizens. Users of the mobile app would be able to create an account, login, report a crime incident, add a location of crime using google map and check the status of crime case.

2) Sequence Diagram (SD): The sequence diagram depicts the flow of events in the system. Fig. 9 shows the sequence diagram (SD) for recording a case.

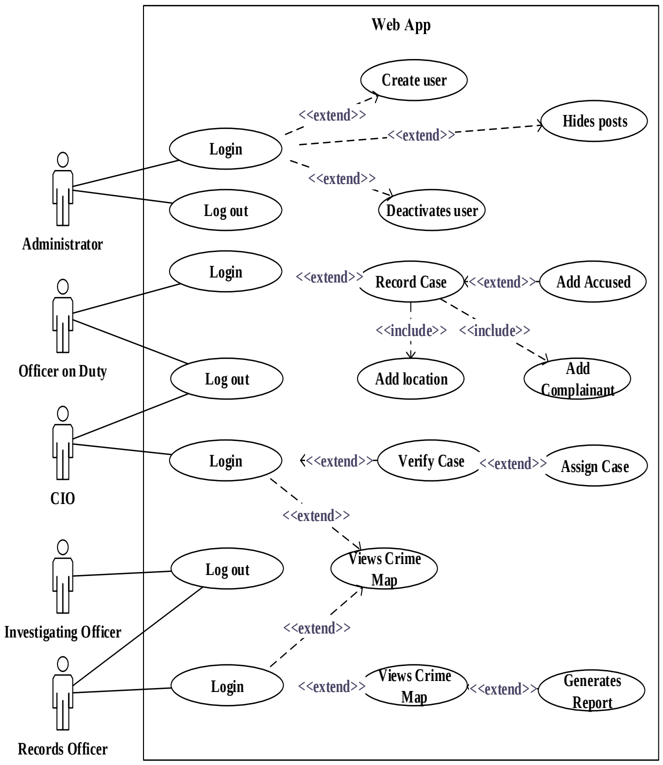


Fig. 7. Use Case - Web App.

3) Entity Relationship Diagram (ERD): The entity relationship diagram (ERD) for the proposed system is shown in Fig. 10.

The ER model diagram in Fig. 10 shows the relationships between entities and attributes in the proposed system.

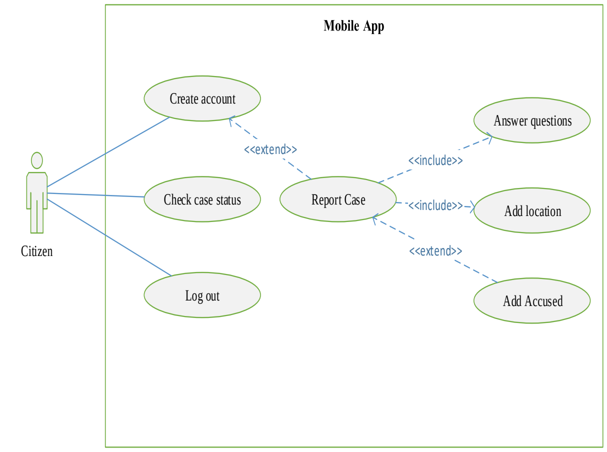


Fig. 8. Use Case – Mobile App.

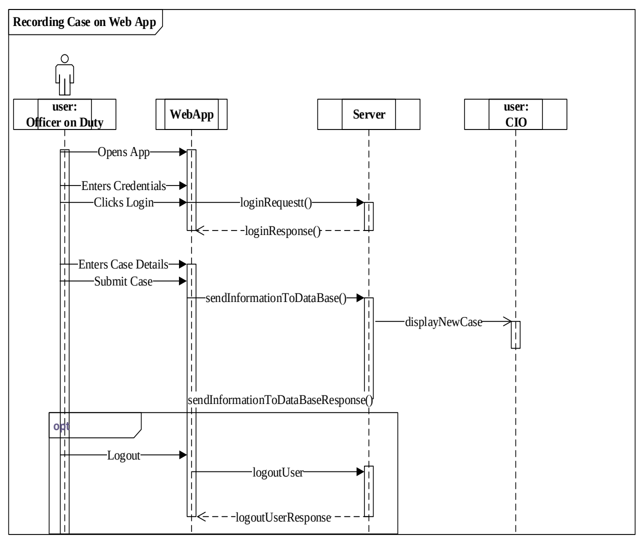


Fig. 9. Sequence Diagram -Case Recording.

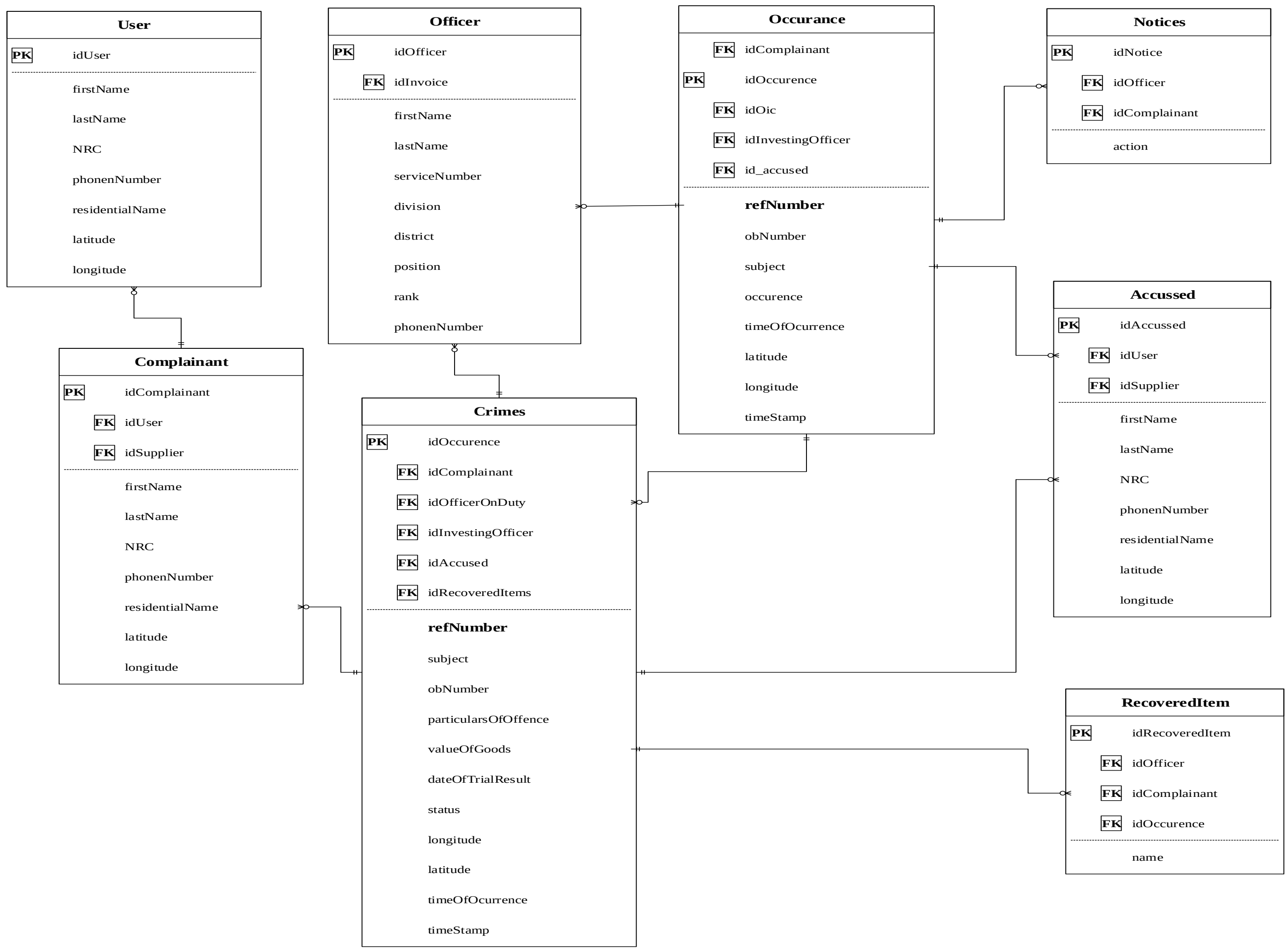


Fig. 10. Entity Relationship Diagram.

IV. Results

The results obtained from the baseline study and system prototype development and testing are presented in this section. The main purpose of conducting the baseline study was to identify the challenges in the current crime mapping system used by the Zambia Police Service. A proposed prototype application was developed to address the challenges.

A. *Baseline Study*

The data collected from the baseline study was analyzed using descriptive statistics and the results were presented in different forms. A study was conducted to find out the levels of knowledge in computers among the officers in the Zambia Police, the results are shown in Fig. 11.

As shown in Fig. 11, all the officers interviewed indicated that they had at least some basic knowledge of computers, with more than 50% of the respondents indicating that their level of knowledge in computers was either good or excellent.

A study was conducted to find out on the usage of crime mapping technique in police stations, the results are shown in Fig. 12.

The results in Fig. 13 show that 85.2% of the stations conduct crime mapping using manual or physical geographical maps and pins placed on the map while 14.8% indicated that they don’t use any form of crime mapping technique.

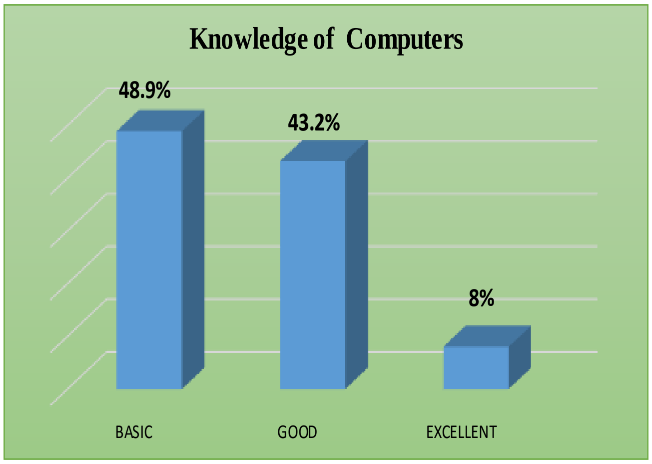


Fig. 11. Computer Knowledge.

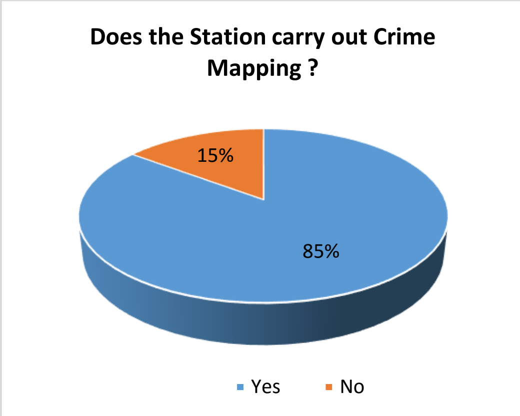


Fig. 12. Crime Mapping usage.

The study revealed that one of the challenges of the current manual system is non-instant availability of statistics or data on areas prone to crimes as indicated by 95% of respondents in Fig. 13.

As shown in Fig. 13, one of the major cause of non-instant availability of crime data is the manual storage of data that requires physical counting. The manual data does not support data querying hence data is subjected to be counted one by one to establish the desired statistics, a process which is known to be time consuming.

A total of 91% of respondents further indicated of having challenges in analyzing & managing crime data with manual maps as shown in Fig. 14.

As shown in Fig. 14, the major weakness of physical crime map is that it does not only support querying techniques but also can only accommodate one particular crime data at time.

Further, the study looked at how crimes are reported by the general public and how the police capture and record crime data, Table I shows the crime reporting methods.

The results in Table I indicate that 75.9% of people walk to the station to report crimes. Only 24.9% make phone calls. In addition, all crime cases and statistics are manually captured, recorded, processed and stored in books & papers. Data input is critical for the accurate of digital maps. For digital crime maps to be automated there is need to consider digitalizing the stage of reporting, capturing and recording of crime data by introducing a mobile application platform to be used by the general public to report crime case. In order to understand the type and nature of mobile devices used by the general public, questionnaires were distributed to citizens who were visiting the police stations for various reasons across Lusaka city. The results in Fig. 15 show that 86% of the general public own phones or other mobile devices that have access to internet.

The study further wished to know and uncover the type of the mobile operating systems of the mobile devices considered in the survey, the results are shown in Fig. 16.

As shown in Fig. 16, for the mobile devices that have internet access displayed in Fig. 15, 72% of those had android operating system,14% had Microsoft windows while only 8% have Apple and 6% had unknown mobile operating systems. With these results it is evident that in order for the system to be accessed by majority citizens it had to be belt on both web and android application platforms.

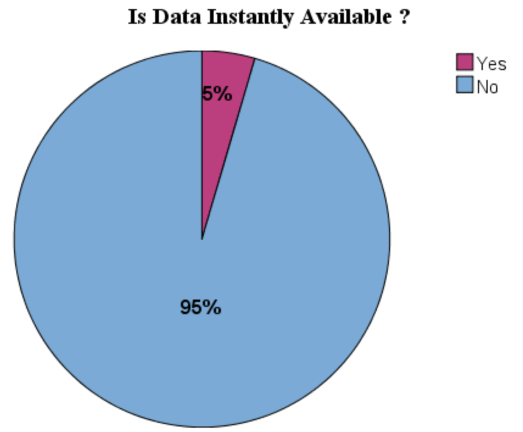


Fig. 13. Non Instant Availability of Data.

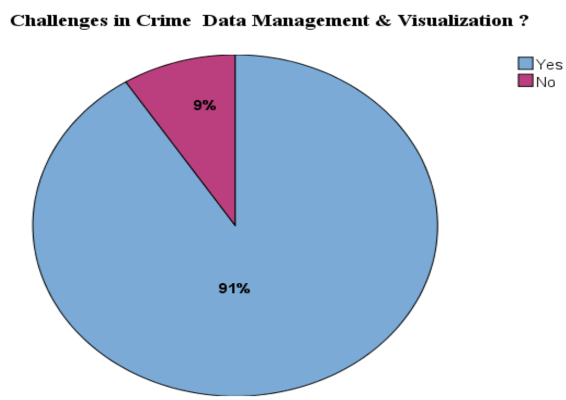


Fig. 14. Challenges in Data Management and Visualization.

TABLE. I. CRIME REPORTING METHODS

How does the public report crimes ?			
	Responses		Cumulative frequency
	Frequency	Percent	
Does the public walk in to the police station when making crime reports?	88	75.9%	100.0 %
Does the public make a call to the police when making crime reports?	28	24.1%	31.8%
Total	116	100.0%	131.8 %

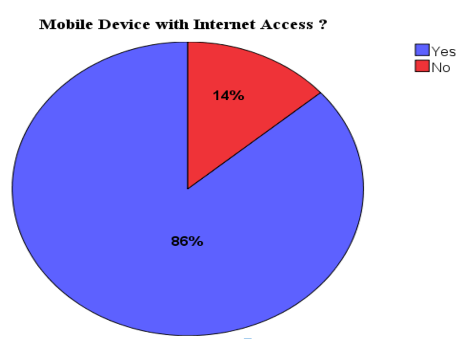


Fig. 15. Mobile Devices.

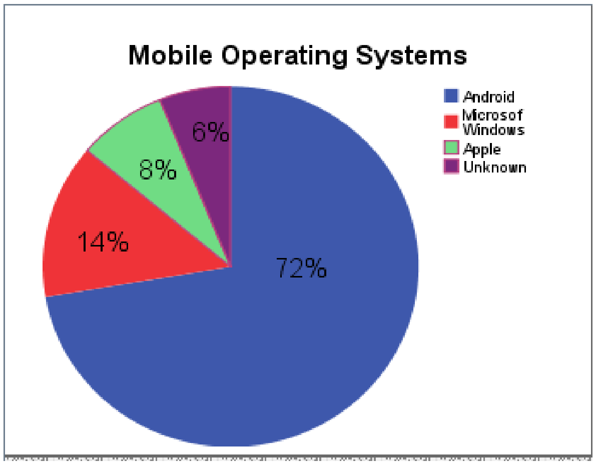


Fig. 16. Mobile Operating System.

B. System Prototype Development

As already outlined in the previous section, the prototype application named crime mapper consists of the web and android mobile platforms. The web application will strictly be used only by the police consisting of front end which is a web browser and a sever backend consisting of firebase web server and cloud firebase database for data storage. The police can use the web application to record crime incidences from the citizens who walks into the police station, generate, view and update crime maps. The android mobile application will be running on the user's mobile device which can be used to report crime incidences. Both the web application and mobile application are connected to one central database in the cloud.

1) Mobile application: The mobile application will only be used by the general public to report crime incidences, the user is required to download the application and register his or her details into the system as shown in Fig. 17.

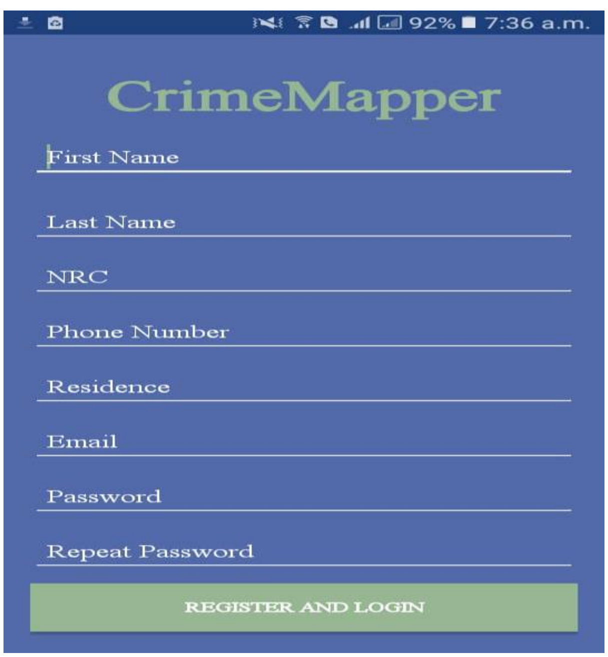


Fig. 17. User Registration – Mobile App.

As shown in Fig. 17, the registration details specified by the user will be saved into the database and displayed in the occurrence book at the police station. The next time the user wants to use the application, he or she will be prompted to sign in as shown in Fig. 18.

Once the user successfully registers or logs into the application, the system will direct the user to the crime incident reporting screen as shown in Fig. 19.

As shown in Fig. 19, the user is required to select and enter the category of crime case for example rape, murder, theft and many others. In addition, the user is required to specify the location time and location of crime incident. The location of crime is selected from the google map shown in Fig. 20.

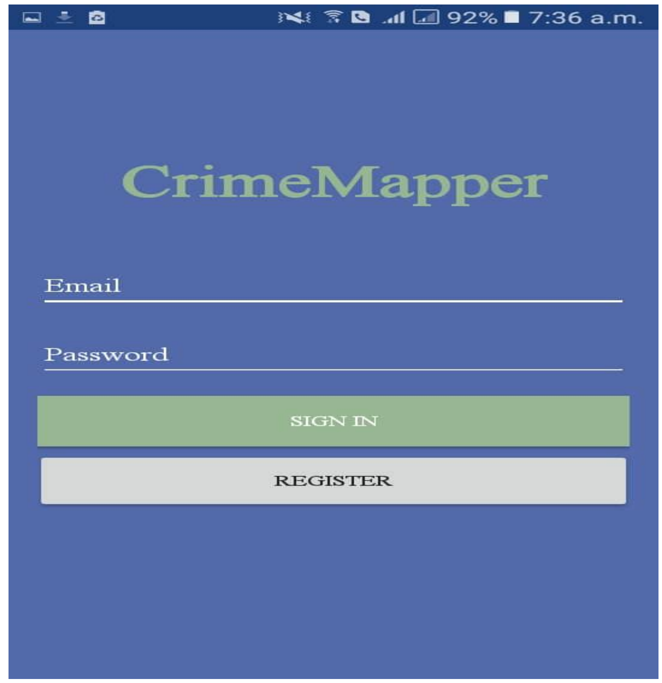


Fig. 18. Login Mobile App.

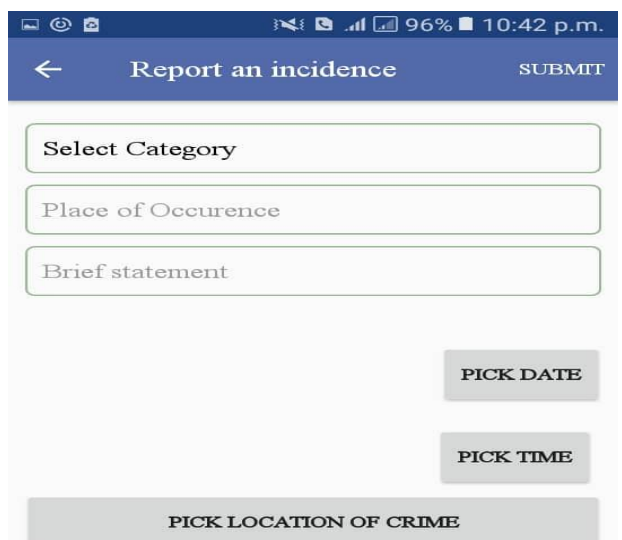


Fig. 19. Crime Reporting – Mobile App.

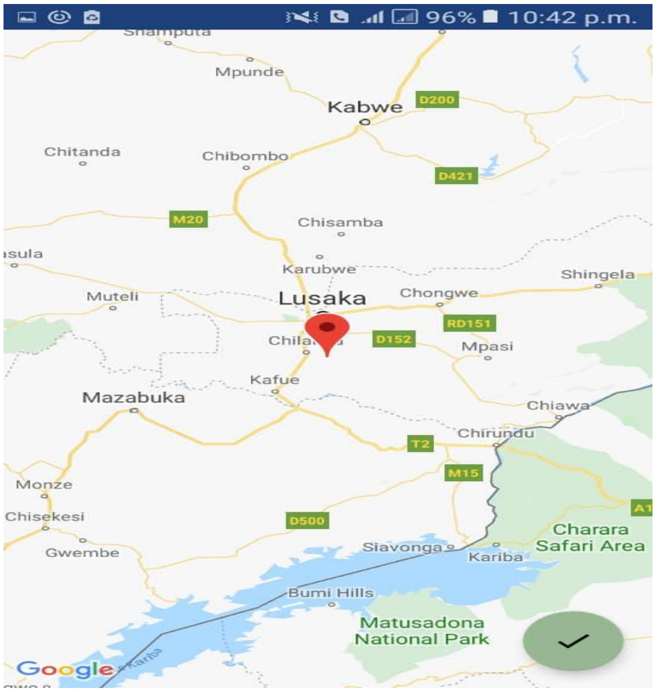


Fig. 20. Crime Location Navigator – Mobile App.

As shown in Fig. 20, the user navigates to the actual location of crime on the map, the crime location coordinated together with other crime details are saved into the occurrence book and crime register designed in the cloud database which can only be accessed by the police using the web application. Based on the crime location coordinates, the crime incident report is directed to a nearby police station. The user is able to view the status of the case they reported as shown in Fig. 21.

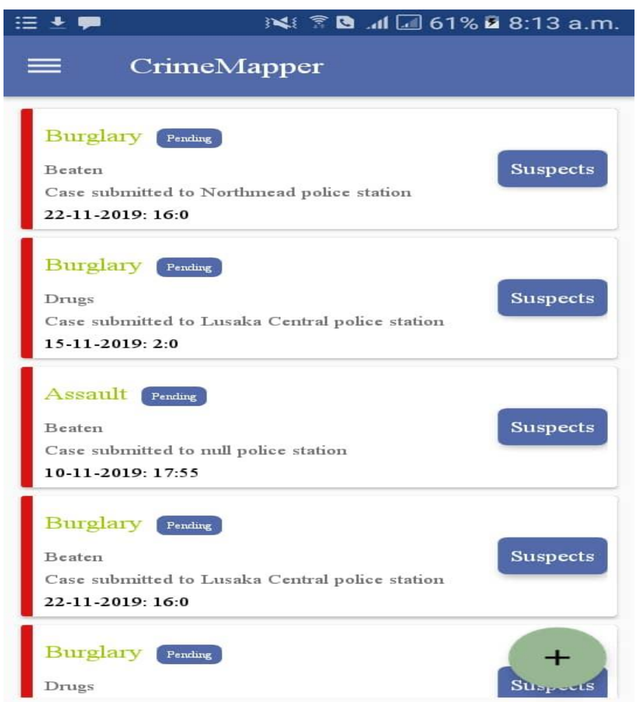


Fig. 21. Crime Status – Mobile App.

2) *Web application:* The web application will only be used by the police to record crime incidences from the general public that walks or physically visits the police station. The application will be used to view and generate crime maps and statistical reports. The users who are the Police Officers are required to register their details into the system as shown in Fig. 22.

If the details of the user are already captured into the application, the system prompts him or her to login as shown in Fig. 23.

When the user successfully logs into the application, the system directs to the window where crime reports are listed as shown in Fig. 24.

The user can record a new case by clicking on add new case as shown in Fig. 25.

The user can specify the location of the crime by navigating on the google map as shown in Fig. 26.

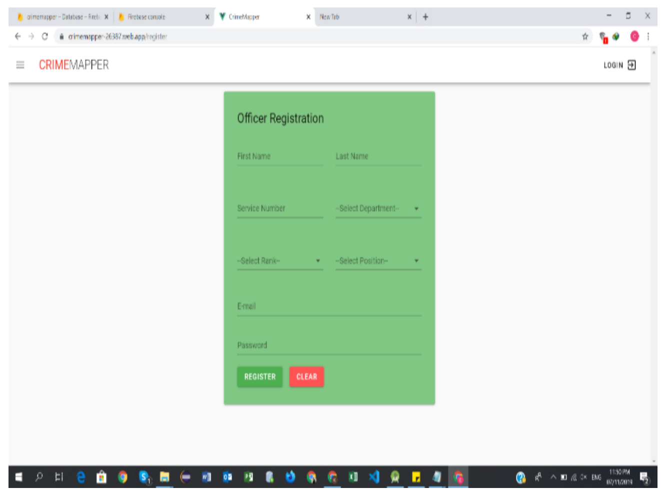


Fig. 22. User Registration – Web App.

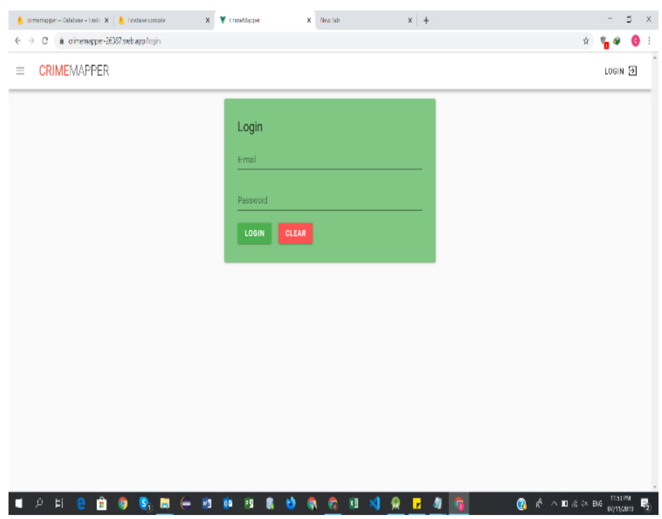


Fig. 23. Login – Web App.

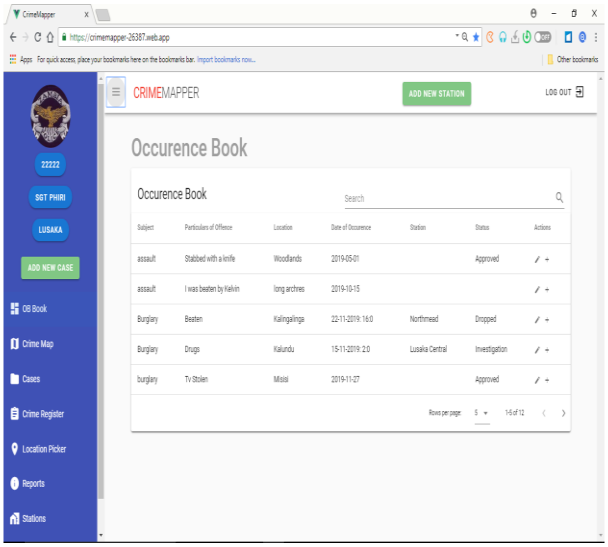


Fig. 24. Crime Reports List – Web App.

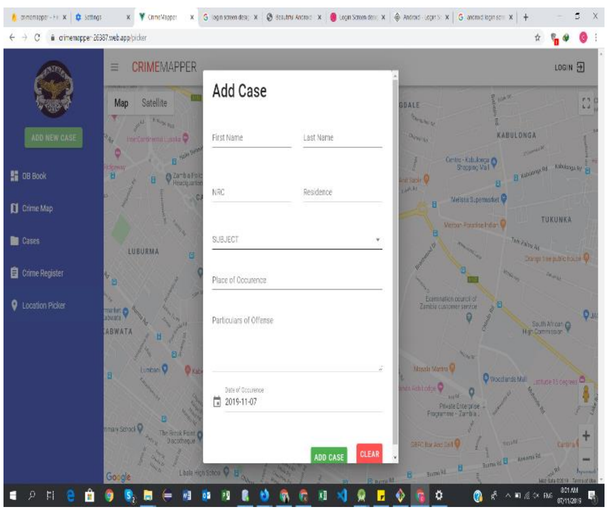


Fig. 25. Adding a Crime Report –Web App.

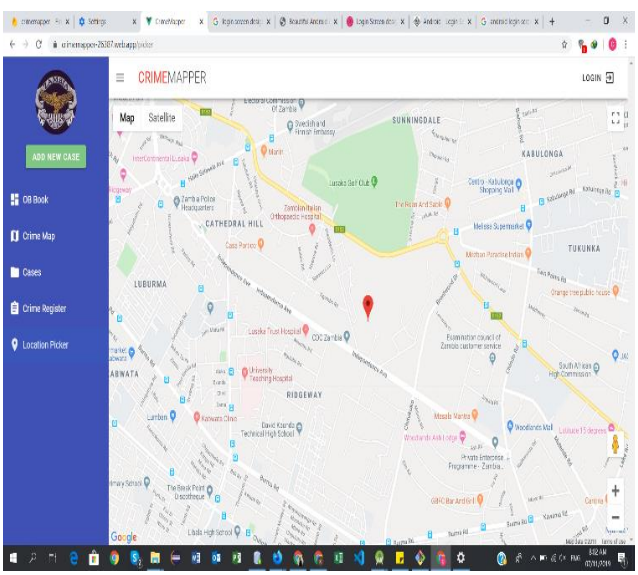


Fig. 26. Crime Location Navigator – Web App.

As shown in Fig. 26, the location navigator enables the user to select the exactly location of a crime from a Google map. The location coordinates are saved into the central database in the cloud. The screenshot window in Fig. 27 shows the filtering of the map based on the name of crime.

As shown in Fig. 27, the crime map can be filtered using the name of a crime like murder, rape, theft and others types of crime. This means a user can choose specific spatial spots to be display on the crime map.

Fig. 28 shows a screen shot window displaying sample spatial location of crime spots.

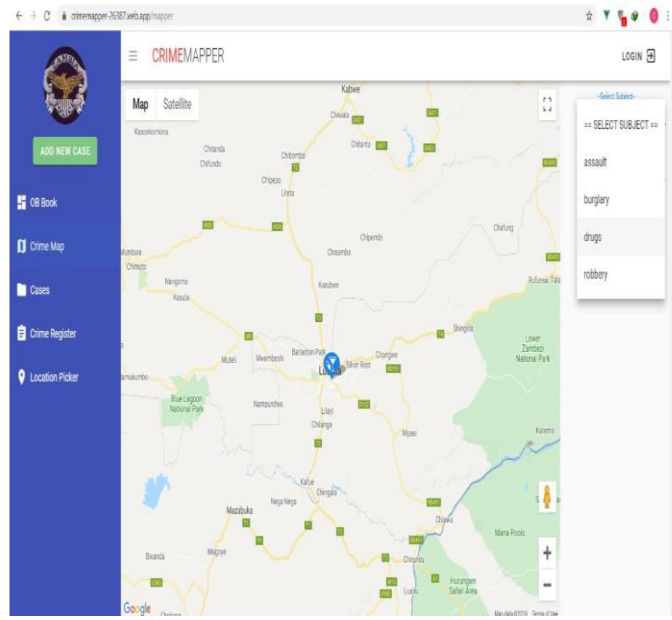


Fig. 27. Filtering Map – Web Map.

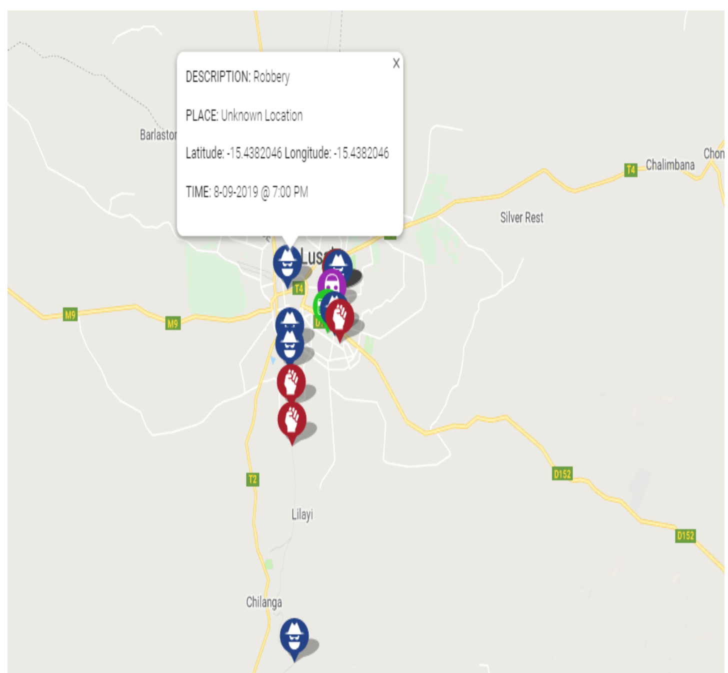


Fig. 28. Spatial Crime Spots – Web App.

As shown in Fig. 28, the system is able to display spatial crime spots as they are being reported by the general public. Whenever a crime is reported and recorded into the system through either mobile application or web application, a pin is placed on the map representing a crime incident. Each crime sport shows name, location and its coordinates, date and time the crime was reported.

V. Discussion

The study was aimed at identifying and establishing challenges of the current crime mapping system used by the Zambia Police service and also design a computerized crime mapping model for reporting and mapping of crimes. From the baseline study conducted, it was discovered that 85.2% of the stations considered in the study conduct crime mapping using manual or physical geographical maps and pins placed on the map while 14.8% indicated that they don't use any form of crime mapping technique. The study revealed that one of the challenges of the current manual system is non-instant availability of statistics or data on areas prone to crimes as indicated by 95% of respondents, the major attribute to this challenge is the manual storage of data that requires physical counting. The manual data does not support data querying hence data is subjected to be counted one by one to establish the desired statistics, a process which is known to be time consuming. The study further showed that total of 91% of respondents further indicated of having challenges in analyzing & managing crime data with manual maps, some of the weakness of physical crime map include not support querying techniques and also can only accommodate one particular crime data at time. The study also looked at how crimes are reported by the general public and how the police capture and record crime data, results indicated that 75.9% of people walk to the station to report crimes. Only 24.9% make phone calls. In addition, all crime cases and statistics are manually captured, recorded, processed and stored in books & papers. For digital crime maps to be automated and implemented there is need to consider digitalizing the stage of reporting, capturing and recording of crime data. The study introduced mobile and web application platforms to be used to capture crime reports. The crime data captured from both the web application and Mobile application should be stored in one central cloud database.

In order to understand the type and nature of mobile devices used by the general public, questionnaires were distributed to citizens who were visiting the police stations for various reasons across Lusaka city. The results show that 86% of the general public own phones or other mobile devices that have access to internet. The study further wished to know and uncover the type of the mobile operating systems of the mobile devices considered in the survey, the results indicated that 72% of those had android operating system,14% had Microsoft windows while only 8% have Apple and 6% had unknown mobile operating systems. With these results it is evident that in order for the system to be accessed by majority citizens it had to be belt on both web and android application platforms.

The current business processes were mapped as indicated in Fig. 3 and a model based on cloud architecture and spatial data was developed as indicated in Fig. 4 and Fig. 5. In order to mitigate the challenges faced by Zambia Police in crime mapping and crime data management, a computerized crime mapping model based on cloud and spatial data was developed, comprising of web application and android mobile application platforms respectively all connected to a central firebase cloud database. The web application will only be used by the Police to not only generate and view crime maps but also capture and record crime reports from the general public. The mobile application will only be used by the general Public to report crime cases to the police. Spatial crime spots are added to the map in real time as crime cases are reported. The test results of the prototype show improved data capture of crime cases and also improved crime data visualization through generated crime maps. Crime maps can be filtered based on the name of crime like murder, theft, rape and other type of crimes. The developed prototype test results also proved that it is more efficient and effective than the current system. For an institution like Zambia Police that is still depending on papers and books for its daily operations, implementing the automated crime mapping system would be a good starting point towards utilization and usage of ICT in Zambia Police Service.

VI. Conclusion

The baseline study was conducted and a number of challenges were identified in the current manual system such as manual capturing and recording of crime reports, non-instant availability of crime data of areas with high crimes, poor crime data management, poor data visualization and many others. The study proposed a computerized crime mapping model based on cloud and spatial data to address the challenges in the current system used by the Zambia Police. The current business processes were mapped and a crime mapping model was designed and developed in order to address the challenges that were discovered in the baseline study. The model is based on the Cloud Architecture, Android Mobile Application, Web Application, Google Map API and Java programming language. Test results of the developed prototype system shows improved spatial crime data visualization and reporting of crime data with reduced dependency on manual transactions, it also proved to be more effective than the current system.

VII. Recommendation and Future Works

A. *Recommendations*

The study has revealed that the automated crime mapping system is desirable. With the police personnel that exhibited high levels of knowledge in computers including a general population that is drifting towards the use of ICT, we recommend the adoption of the proposed model as it will improve the work efficiency within the Police Service.

B. *Future Works*

Some future works that can be done on this system include;

1) The crime dataset generated through crime reporting platforms in this system can be integrated with machine and deep learning techniques to predict future crime occurrences crime hotspot areas.

2) Introducing case docket management would add value to the system. A case docket contains information like address and date of crime, statements, reports from experts, witnesses, details of complainants and suspects. The public prosecutor makes use all the information in the case docket to present a case in the court of law.

ACKNOWLEDGMENT

We would like to thank the Zambia Police Service for the support and also for providing us with all relevant information towards this project. Special thanks goes to UNZA Department of Computer Science and Department of Electrical and Electronic Engineering for the advice and expertise rendered to this project.

REFERENCES

[1] S. Anurag and S. Baghel, "A Predictive Model for Mapping Crime using Big Data Analytics," International Journal of Research in Engineering and Technology, vol. 4, no. 4, pp. 343-348, 2015.

[2] H. Kumar, B. Sainia, and G. Mahajana, "Crime Prediction & Monitoring Framework Based on Spatial Analysis," in International Conference on Computational Intelligence and Data Science, Jaipur, 2018.

[3] V. Kappeler, "What is the Place of Police within the Criminal Justice System?" Eastern Kentucky University, October 2012. [Online]. Available: http://plsonline.eku.edu/insidelook/what-place-police-within-criminal-justice-system. [Accessed 5 Novermber 2016].

[4] Correctionalofficer.org, " U.S Criminal Justice System," correctionalofficer.org, [Online]. Available: http://www.correctional officer.org/us-criminal-justice-system. [Accessed 13 November 2016].

[5] Department of Justice and Equality, "White paper on crime-The Community and the Criminal Justice System," Department of Justice and Equality, Dublin, 2011.

[6] Zambia Police, "Zambia Police History," Zambia Police,2008.[Online]. Available: https://www.google.co.zm/?gws_rd=cr&ei=Pb0aWL3JJciUsg G3ka34Ag#q=history+of+zambia+police+service. [Accessed 29 October 2016].

[7] C.Chandrasekar and M.Kuma, "GIS Technologies in Crime Analysis and Crime mapping," International Journal of Soft Computing and Engineering (IJSCE), vol. I, no. 5, pp. 115-122, 2011.

[8] D. Wang and D. Wei, "Crime hotspot mapping using the crime related factors—a spatial data mining approach," Springer Science+Business Media , New York , 2012.

[9] P. Kedia, "Crime Mapping and Analysis using GIS," International Institute of Information Technology, vol. 1, no. 1, pp. 1-15, 2016.

[10] G. Breetzke, "Geographical Information System (GIS) And Policing In South Africa- A Reveiw," Policing: An International Journal of Policing Strategies and Management, vol. 29, no. 4, pp. 723-740, 2007.

[11] G. Lamond, "What is a Crime?, " Oxford Journal of Legal Studies, vol. 27, no. 4, pp. 609-632, 2007.

[12] K. Harries, Mapping Crime:Principle and Practice, Washington: National Institute of Justice, 1999.

[13] M. Gluck and L. Yu, "Geographic Information Systems," Advances in Librarianship, vol. 1, no. 1, pp. 1-38, 2015.

[14] K. Hansgen, "Crime Mapping In Law Enforcement: Identifying Analytical Tools, Methods, And Outputs," CULMINATING PROJECTS IN CRIMINAL JUSTICE, vol. 1, no. 1, pp. 1-248, 20 October 2016.

[15] C. W. Bruce, "Library-CrimeAnalysis," 3 August 2013. [Online]. Available: http://www.justiceacademy.org/iShare/Library-CrimeAnalysis /CrimeMapping&Analysis.pdf. [Accessed 28 October 2019].

[16] U. Argun and M. Dağlar, "Crime Mapping and Geographical Information Systems in Crime Analysis," International Journal of Human Sciences, vol. 13, no. 1, pp. 1-14, 2016.

[17] A. Murray and J. Western, "Exploratory Spatial Data Analysis Technologies For Examining Urban Crime.," Center for Crime and Justice Studies, vol. 1, no. 1, pp. 1-21, 2001.

[18] S. Chun, V.Avinash and S. Yuan, "Crime Prediction Model using Deep Neural Networks," in Annual International Conference on Digital Government Research, Dubai, 2019.

[19] N. Waduge, "Machgine Learning Approaches for Detecting Crimes," Researchgate, vol. 1, no. 1, pp. 1-7, 2017.

[20] K. Das, R. Behera, "A Survey on Machine Learning: Concept Algorithms and Applications," International Journal of Innovative Research in Computer and Communication Engineering, vol. 5, no. 2, pp. 1301-1309, 2017.

[21] A.Simon, M.Singh ,S. Venkatesan and R. Babu, "An Overview of Machine Learning and its Applications," International Journal of Electrical Sciences & Engineering, vol. 1, no. 1, pp. 22-24, 2015.

[22] A. Dey, "Machine Learning Algorithms: A Review," International Journal of Computer Science and Information Technologies, vol. 7, no. 3, pp. 1174-1179, 2016.

[23] S. Das,A.Dey,A.Pal and N. Roy " Applications of Artificial Intelligence in Machine Learning: Review and Prospect," International Journal of Computer Applications, vol. 115, no. 9, pp. 975-987, 2015.

[24] H.Woo and H.Kang, "Prediction of crime occurrence from multi-modal data using deep learning," 24 April 2017. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.017624 4. [Accessed 10 November 2019].

[25] A.Simon, M. Singh ,S. Venkatesan and R. Babu, "An Overview of Machine Learning and its Applications," International Journal of Electrical Sciences & Engineering , vol. 1, no. 1, pp. 22-24, 2015.

[26] Ben. Benuwa, Y. Zhan , B.Ghansah,D. Keddy and F. Kataka, "A Review of Deep Machine Learning," International Journal of Engineering Research in Africa, vol. 24, no. 1, pp. 124-136, 2016.

[27] I. Namatēvs, "Deep Convolutional Neural Networks: Structure Feature Extraction and Training," Information Technology and Management Science, vol. 20, no. 1, p. 40–47, 2017.

[28] X. Li and Z. Liu, "A Spatial Data Security Model Under the Cloud Environment," Proceedings of the 2nd International Conference On Systems Engineering and Modeling, vol. 1, no. 1, pp. 1-4, 2013.

[29] K. Kavitha, "Study on Cloud Computing Model and its Benefits, Challenges," International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, no. 1, pp. 1-9, 2014.

[30] M. Ahmad and R. Khan, "The Cloud Computing: A Systematic Review," International Journal of Innovative Research in Computer and Communication Engineering, vol. 5, no. 3, pp. 66-75, 2015.

[31] TutorialsPoint, "Cloud Computing Architecture," tutorialspoint.com, 2006. [Online]. Available: https://www.tutorialspoint.com/cloud_ computing/cloud_computing_architecture.htm. [Accessed 20 November 2019].

[32] S. Q. Yang, "Move into the Cloud, shall we?" Library Hi Tech News, vol. 26, no. 1, pp. 4-7, 2012.

[33] R. M. Willet, "Cloud computing assurance – A review of literature guidance," Information and Computer Security, vol. 26, no. 1, pp. 2-33, 2017.

[34] A.Shashikumar and M. Chandrashekhar , "Cloud Computing Services and Deployment Models," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 5, no. 6, pp. 1083-1088, 2017.

[35] C. Rao and R. Kumar, "Cloud Computing Services And Deployment Models," International Journal Of Engineering And Computer Science, vol. 2, no. 12, pp. 3389-3392, 2013.

[36] T. Diaby and B. Rad, "Cloud Computing: A review of the Concepts and Deployment Models," Information Technology and Computer Science, vol. 6, no. 1, pp. 50-58, 2017.

[37] G. Huilian and X. Wang, "Cloud-based Service for Big Spatial Data Technology in Emergence Management," Henan Polytechnic University, jiaozuo, 2001.

[38] N. Mwansa and J. Phiri, "Real-Time Data Capture Model for Accelerated Payment of Small-Scale Farmers," American Journal of Engineering and Applied Sciences, vol. 11, no. 3, pp. 1164-1177, 2018.

[39] C. Mamalian & N.LaVigne, "The Use of Computerized Crime Mapping by Law Enforcement: Survey Results," USA National Institute of Justice, Washington, 1999.

[40] P. Chamikara & J. Gunathilake, "Intelligent Crime Mapping System," University of Peradeniya, KAndy, 2012.

[41] G. D. Breetzke, "Geographical Information Systems (GIS) and Policing in South Africa- A Review," An International Journal of Policing Strategies and, vol. I, no. 3, pp. 1-82, 2007.

[42] H. Tong, "A Crowdsourcing Based Crime Mapping System," Waseda University, Tokyo, 2014.

[43] H. Singh, R. Kumar, A. Singh, P.K. Litoria, "Cloud GIS for Crime Mapping," International Journal of Research in Computer Science, vol. 2, no. 3, pp. 57-60, 2012.

[44] D. Summerton, D. Mazeika, "The impact of geocoding method on the positional accuracy of residential burglaries reported to police," Policing: An International Journal of Police Strategies & Management, vol. 40, no. 2, pp. 2-15, 2017.

[45] O.Hutt, K. Bowers, S.Johnson and T. Davies, "Data and evidence challenges facing place-based policing," Policing: An International Journal of Police Strategies & Management, vol. 41, no. 3, pp. 339-351, 2018.

[46] M. Mwiya, J. Phiri and G. Lyoko, "Public Crime Reporting and Monitoring System Model Using GSM and GIS Technologies: A Case of Zambia Police Service," International Journal of Computer Science and Mobile Computing, vol. 4, no. 11, pp. 207-226, 2015.

Classification Models for Determining Types of Academic Risk and Predicting Dropout in University Students

Norka Bedregal-Alpaca[1], Víctor Cornejo-Aparicio[2], Joshua Zárate-Valderrama[3], Pedro Yanque-Churo[4]
Universidad Nacional de San Agustín de Arequipa, Arequipa, Perú

***Abstract*—Academic performance is a topic studied not only to identify those students who could drop out of their studies, but also to classify them according to the type of academic risk they could find themselves. An application has been implemented that uses academic information provided by the university and generates classification models from three different algorithms: artificial neural networks, ID3 and C4.5. The models created use a set of variables and criteria for their construction and can be used to classify student desertion and more specifically to predict their type of academic risk. The performance of these models was compared to define the one that provided the best results and that will serve to make the classification of students. Decision tree algorithms, C4.5 and ID3, presented better measurements with respect to the artificial neural network. The tree generated using the C4.5 algorithm presented the best performance metrics with correctness, accuracy, and sensitivity equal to 0.83, 0.87, and 0.90 respectively. As a result of the classification to determine student desertion it was concluded, according to the model generated using the C4.5 algorithm, that the ratio of credits approved by a student to the credits that he should have taken is the variable more significant. The classification, depending on the type of academic risk, generated a tree model indicating that the number of abandoned subjects is the most significant variable. The admission scan modality through which the student entered the university did not turn out to be significant, as it does not appear in the generated decision tree.**

Keywords*—Educational data mining; ID3 algorithm; C4.5 algorithm; artificial neural network; classification algorithms; student desertion; academic risk

I. Introduction

Education is fundamental to the development and well-being of a society, so students are the fundamental asset of any educational institution. The social and economic development of a country is directly related to the academic performance of its students [1]. In this context and because of the implications it has, university desertion is a problem that raises concerns in the managers of higher education institutions; on the one hand, the university's finances are affected and on the other hand, the efficiency of the higher education system is questioned, because only a small number of young people who start university studies manages to complete them.

One way to address this problem is to have timely information about the possibility of student dropout, hence the interest in analyzing the possible factors that may lead a student to enter conditions of dropout. Today, there are multiple applications from artificial intelligence to education, data mining techniques are used to discover important patterns and obtain useful information from academic-based information systems [2]. These techniques are known as Data Mining, which analyzes and understands student-related information; so it can transform this information into a useful way that can improve an institution's decision-making to improve the quality of education.

In [3] it is used artificial neural networks, specifically with a Multilayer Perceptron topology, to predict student performance. The attributes selected are mainly of two types, first academic attributes related to the academic details of the students and the personal attributes among them are: study interest, unit test notes, assignment, extracurricular activities, residency, parent education and family status. As a result, 91% accuracy was achieved with MLP training.

In [4] the ID3, J48, Bayes Net, and Naive Bayes algorithms are applied to a sample of 577 students to predict their performance on the first semester exam. Comparative analysis of the results of applying these algorithms has helped students improve and focus more on different courses.

The work of [5] focuses on the development of mining models to predict student performance, based on personal, pre-university and university performance characteristics. Well-known classification algorithms were applied to the data set including a "rule learner", a decision tree classifier, a neural network, and the nearest neighbor classifier. Rated accuracy is obtained between 67.46% and 73.59%. The highest accuracy is obtained for the neural network model (73.59%), followed by the decision tree model (72.74%) and the k-NN model (70.49%).

In [6], the C4.5 decision tree algorithm is used to apply it to students' internal assessment data to predict their performance on the final exam. The decision tree result predicted the number of students who are likely to fail or approve. It is obtained that the C4.5 algorithm is more efficient than ID3 in terms of accuracy and the time needed to train the tree and build it.

In [7] it is conducted an experiment on eight actual datasets using five C4.5 algorithm methods based on entropy from Shannon, Havrda - Charvt, Quadratic, R'enyi and Taneja, making a comparison between them and building a model that takes the entropies of Shanon, quadratic and Havrda-Charvt in parallel and produce more accurate classifications for the dataset and a result of this classification is comparable to the other machine learning techniques. This approach based on entropies can be applied to real-world classification problems.

This work focuses on classification techniques, techniques that use a set of pre-classified examples to develop a model that can classify a population of similar records [8]). Classification techniques are applied under two different approaches: the first related to student desertion which will be represented by two classes: the student ABANDON the career and the student DON'T ABANDON the career; and the second that identifies the type of risk of students from those who in the first approach the resulting class was that they were leaving the race; This is intended to identify whether the student presents an INCIPIENT, MODERATE, HIGH or VERY HIGH risk of dropping out of university studies.

It makes use of an own-build application where the functionality of creating classification models for different algorithms such as artificial neural networks and decision trees was implemented; of the latter specifically used the algorithm ID3 and C4.5. An analysis of the variables chosen as inputs to the classification algorithms will be performed and the performance metrics of the models generated for each of the proposed approaches will be compared. Subsequently the best models can be used to classify the desertion or risk type of future students.

II. Theoretical Mark

A. Classification

It is the most commonly applied data extraction technique; it can predict the value of a categorical attribute (discrete or nominal). Uses a set of preclassified examples to develop a model that can classify the overall record population [9]; uses a decision tree or neural network-based classification algorithms. The process involves two phases, learning and classification. In the learning phase, the classification algorithm analyzes the training data. In the classification phase, test data is used to estimate the accuracy of the classification rules. If accuracy is acceptable, rules can be applied to new data tuples. The classifier training algorithm uses the preclassified examples to determine the set of parameters required for proper discrimination, and then encodes those parameters into a model called classifier.

B. Artificial Neural Networks (ANN)

They are computing algorithms that can solve complex problems by mimicking animal brain processes in a simplified way [10]. They are based on a directed graph structure, composed of a set of neurons that interconnect through arcs directed with an associated weight that determines the force and sign of the connection. Neurons are organized by levels or layers and have two defined functions: activation and output. Their main advantage is that they can solve problems that are very complex for conventional technologies; problems that do not have an algorithmic solution or for which an algorithmic solution is very complex to be defined. Another advantage is that they are very robust with respect to noise in the data.

The best-known ANN is Multilayer Perceptron (MP), which comprises an input layer, a hidden layer, and an output layer, this type of network is being used to perform data mining processes [11]. MP networks consist of neurons arranged in layers and interconnected by synaptic weights and can filter and transmit information, in a supervised manner, to build a predictive model that classify data stored in memory.

C. Decision Trees

They are tree-shaped structures whose nodes represent a choice between several alternatives, and each leaf node represents a decision [4], [8]. They are generally used for classification, because they are simple hierarchical structures that facilitate user understanding. The training algorithms are simple and their computation time is short, as they only require traversing the tree until they reach the leaf node to perform the classification. They use real data extraction algorithms to help with classification. They are used as support for the choice between various lines of action, allowing to explore the possible results for various options, and evaluate the risk and rewards for each possible course of action. These decisions generate rules, which are then used to classify the data. Decision tree learning algorithms include ID3, C4.5 and ASSISTANT.

D. Algorithm ID3 – Iterative Dichotomizer 3

It was invented by Ross Quinlan, iteratively divides the attributes into two groups: the most dominant and the others to build the tree from top to bottom. ID3 uses information theory to determine the most informative attribute, for this there are two concepts involved: entropy and information gain. If the entropy of the attribute is zero, it is a homogeneous node and does not need to be classified; if your entropy is 1, it is a heterogeneous node and you need to continue to classify it. Built the tree, it is applied to a tuple in the database, resulting in the classification of that tuple. The sample dataset must consist of a series of tuples of values, each called attributes, in which one of them (the attribute to classify) is the target, which is of binary type (positive or negative, yes or no, valid or invalid).

Applies to discrete attributes, at each node an attribute is selected and a value is selected, for example, Average - High. The ID3 algorithm never produces trees that are too large, which makes it easier for the user to read and interpret it.

E. Algorithm C4.5

Based on Hunt's C4.5 algorithm, it handles categorical and continuous attributes to build a decision tree. The C4.5 algorithm is proposed in [12] as an improvement of the ID3 algorithm, which eliminates many of its limitations, for example:

- Allows the possibility to use continuous data by separating the possible results into two branches according to a selected threshold, so that all values above the threshold are assigned to one son and the rest to the other son. For example, Average >15 and Average <-15.
- Allows you to use attributes with missing values, where the example of unknown value is given the value that appears most in the other examples.
- Has the ability to use attributes with different weights.

III. Materials and Technicals

A. Materials

The data correspond to the academic records of 970 students belonging to the Professional School of Systems Engineering. The input variables considered for the study are described in Table I.

The output variables or classes in which a student will be classified are: for the first approach ABANDON AND NOT ABANDON and for the second approach are INCIPIENT RISK, MODERATE RISK, HIGH RISK and VERY HIGH RISK.

The application used to implement, execute and compare algorithms and their results was developed using the Java programming language, the MySQL database engine. JPA was also used, it is a useful Java framework for handling relational data.

B. Data Mining Techniques

In this research, the data mining techniques proposed for classifying students and predicting the risk of desertion (Abandon, Don't Abandon) are: artificial neural networks, C4.5 algorithm and ID3 algorithm.

TABLE. I. Variables and their Values

Variable	Data type	Value
Gender	Nominal	0 – Male 1 – Femaleo
Average Grades of Career Subjects	Numeric and Nominal	0 a 1 (Normalized with min=0 y max=20) Very bad (0 - 7), Bad (7-11), Regular (11-14), Good (14-16), Very good (16-20)
Average Grades General Subject	Numeric and Nominal	0 a 1 (Normalized with min=0 y max=20) Very bad (0 - 7), Bad (7-11), Regular (11-14), Good (14-16), Very good (16-20)
School of Origin	Nominal	1 - State 2 - Particular 3 - Parochial
Career Subject Credit Ratio	Numeric and Nominal	0 a 1
General Subject Credit Ratio	Numeric and Nominal	0 a 1
Admission test score	Numeric and Nominal	0 a 1 (Normalized with min=0 y max=100) Regular (0-50), Good (50-75), Very Good (75-100)
Admission type	Nominal	1 – Ordinary 2 – Extraordinary
The credit relationship the student has with respect to those he should have	Numeric and Nominal	0 - 1
Number of Abandoned Subjects	Numeric and Nominal	None, One, Two, More than two
Number of Disapproved	Numeric and Nominal	None, One, Two, More than two

C. Techniques for Evaluating Classifiers

The evaluation of classification techniques is important, as it allows to validate the goodness of fit of the model in relation to the training set. In addition, the evaluation allows to compare between several classification techniques and select the one that provides the most precision. For the evaluation of the data mining techniques used in this work, 3 performance measures were calculated: accuracy, precision and sensitivity.

IV. The Proposal

A. Knowledge Discovery in Databases (KDD)

For the development of this proposal was followed the process of Knowledge Discovery in databases that allows to identify valid and potentially useful patterns to generate knowledge and make decisions from it. The steps in this process are shown in Fig. 1.

B. Data Selection and Pre-Processing

It was used as input the data of students from the Professional School of Systems Engineering of the National University of San Agustín (UNSA) in Arequipa, Peru. The information was provided by the UNSA Institute of Computing, using 3 files with an .xlsx extension. Archives contained:

- Basic information of students such as: first and last names, C.U.I. (unique identification code), date of birth, gender, and type and place of the school of origin.
- Record of final grades of students in the subjects considered in the curriculum. It is considered the name and code of the subject that a student took (identified by his C.U.I.), the group, the enrollment number, the grade, the condition at the end of the course (approved, disapproved or abandoned the subject), the year and academic cycle in which the student took the subject.
- Score obtained by the student in the entrance test, the type of admission and the position in which he/she entered.

The data in the training set must be pre-processed before being evaluated by the algorithms. Initial data were for 970 students, data cleansing was performed to eliminate noise and isolated data or some inconsistencies. For example, there were students who did not have information about the entrance exam or lacked information about the school of origin or the subjects in which they were registered. Those cases were removed from the final dataset. Some typographical or formatting errors were also fixed. The pre-processed data set consisted of 451 students, of which 338 students were taken, representing 75% to be able to train and generate the ranking models. Of these 338 students, there were 215 instances with the class "DO NOT ABANDON" and 123 instances with the class "ABANDON". The remaining 25%, which is equivalent to 113 students were used for the trial phase, where 88 of these instances have the "DO NOT ABANDON" class and 25 with the "ABANDON" class.

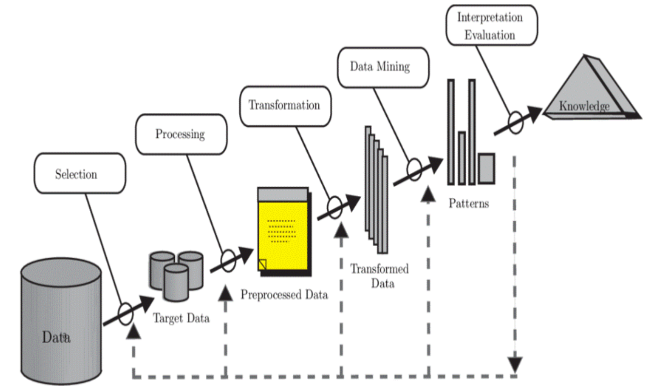


Fig. 1. Steps to Extract Knowledge from the Data (Source: Baradwaj [12]).

For the second approach, a training and testing set consisting of those instances of the previous set of 338 students where the already known class is "ABANDON" was considered to determine the type of risk for each: INCIPIENT RISK, MODERATE RISK, HIGH RISK AND VERY HIGH RISK. Because of this, a set of 160 student instances was used for this approach, 137 belong to the training set and 23 to the test set. Fig. 2 shows a part of this last set (the information appears in Spanish, as it is the language in which the information was worked on).

One point to keep in mind when splitting datasets is that it has enough information about each of the input variables and their classes but that it's also not too much not to over-adjust the training algorithm so it requires to be found adequate proportions. Fig. 3 and 4 show statistical graphs on the percentage of occurrence of each class of a variable in the set and the number of students belonging to a variable class, respectively.

C. Data Transformation

Model training requires inputs to the values of the variables corresponding to each of the students within the training set. The average grades obtained by a student is one of the most commonly used variables in other research around academic performance. Other variables were also considered: the number of credits that the student should possess at the end of a semester, number of subjects disapproved or abandoned during the semester, type of school of origin, gender and notes related to his admission to college.

	A	B	C
1	CUI	CLASE	TIPO
2	20110189	Riesgo moderado	ENTRENAMIENTO
3	20110343	Riesgo moderado	ENTRENAMIENTO
4	20110589	Riesgo incipiente	ENTRENAMIENTO
5	20110729	Riesgo incipiente	ENTRENAMIENTO
6	20111426	Riesgo bajo	ENTRENAMIENTO
7	20111427	Riesgo incipiente	ENTRENAMIENTO
8	20111428	Riesgo incipiente	ENTRENAMIENTO
9	20111429	Riesgo moderado	ENTRENAMIENTO
10	20111430	Riesgo moderado	ENTRENAMIENTO
11	20111431	Riesgo incipiente	ENTRENAMIENTO
12	20111432	Riesgo incipiente	ENTRENAMIENTO
13	20111433	Riesgo moderado	ENTRENAMIENTO
14	20111434	Riesgo moderado	ENTRENAMIENTO

Fig. 2. Data File Provided to the System (Student Representation, Exit Class, and Set Type).

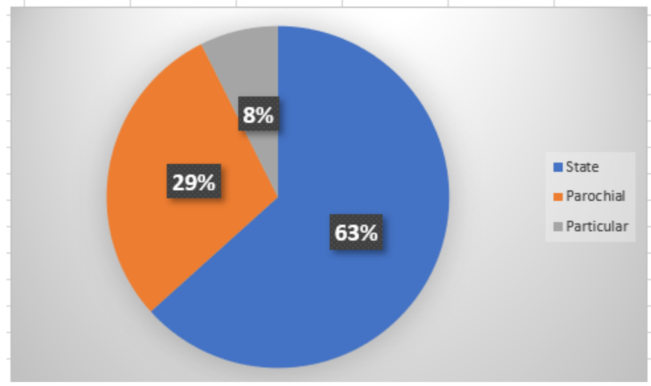


Fig. 3. Data Sharing for the Variable "Scool of Origin" in the Training Set.

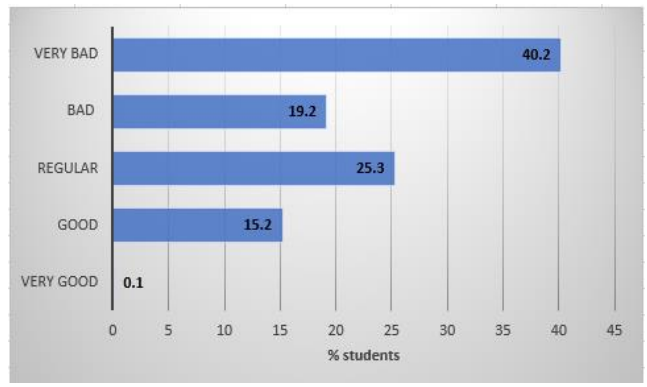


Fig. 4. Data Sharing for the "Califications Average" in Training Set.

To generate classification models for each of the implemented algorithms (neural networks, ID3 and C4.5) the values of the mentioned variables must be provided as inputs, however; each algorithm receives these values differently so it is necessary to transform these values to an appropriate representation.

For example, in the case of the ID3 algorithm, it can only work with nominal attributes, i.e. if working with the gender variable, the values 'FEMALE' and 'MALE' would be used. On the other hand, the C4.5 algorithm supports both nominal and continuous data, i.e. for the case of the average note variable could work with the continuous values which in this case would be a decimal value between 0 and 20; or use nominal or categorical values as very good (16 to 20), good (14 to 16), regular (11 to 13), bad (7 to 10), very bad (0 to 6). For neural networks it is necessary to normalize the data to values between 0 and 1 due to the activation function used, sigmoidal, which has a range of values between 0 and 1.

D. Training Phase

After having the transformed data, the models are trained. The system was provided with a file with the input variables (gender, average notes, credit ratio, etc.) and a field classified as output (abandon or not) of the model. It was also provided with a file listing the student codes that made up the training set along with the class tag to which they belonged, a classification that was previously made based on real information. This action is justified in the training being performed using known data and when the actual classification is performed, it is done on unworked data either in the training phase or in the test phase, so the results thus obtained are a prediction.

A matrix of Student/Variables is then constructed which contains, for each student within the dataset, the value of each of the selected variables (Table II). The Student/Variables matrix serves as input to the classification algorithm that will initiate your training.

If the chosen option is a neural network, a multilayer perceptron network trained with the backpropagation algorithm is built. To stop the training, two different criteria were established: reach an error rate of 0.01% or reach a maximum of 100000 iterations. At the end of the training, the weights of the connections of the neurons were saved in a file, which would allow the network to be reassembled in case the model wanted to be used again.

In the case of the ID3 algorithm, the entropy and information gain of each attribute is calculated, and in this way the most dominant attribute can be found. The most dominant attribute is placed in the tree as a decision node. After that, entropy and profit scores would be calculated again among the other attributes. Thus, the next most dominant attribute is found. This procedure continues until a decision is reached for that branch. Once it is built tree, it is applied for a tuple in the database and this results in the classification for that tuple.

For C4.5 training is very similar to ID3, but in this case the gain ratio of each attribute is calculated.

The output variables will depend on the approach used where they will be “ABANDON” and “DO NOT ABANDON” if try to determine the desertion of students; whereas it will be classified into the four different types of risk if the second approach is used.

After the training phase, the respective tests must be performed with the test set. For each trained model, 3 performance measures were calculated: accuracy, precision and sensitivity.

To perform the training phase, the system was provided with the set of records that belonged to the test suite. Similarly to the training phase, the class to which the student belongs was specified the goal was to apply the classification model to each record and compare the output with the actual class to which the record belongs and apply the performance measures as it is shown in Fig. 5, where in the Student/Variable matrix, it has a column indicating the actual class to which the student belongs and another column indicating the classification made by the model used. Performance metrics are subsequently generated.

TABLE. II. STUDENT/VARIABLE MATRIX

Student	**Variable**			
	V1	**V2**	**…**	**Vm**
E1	E1V1	E1V2	…	E1Vm
E2	E2V1	E2V2	…	E2Vm
E3	E3V1	E3V2	…	E3Vm
…	…	…	…	…
En	EnV1	EnV2	…	EnVm

CUI	GENDER	CAL AVERAGE	CREDITS RELATION	SCHOOL ORIGIN	ADMISIO N SCCORE	CLASIFICA TION	RESULT
20110189	FEMALE	12.0	92.0	PARTIC	65.0	NO ABAN	NO ABAN
20110343	MALE	12.0	80.0	PARTIC	68.0	NO ABAN	NO ABAN
20110689	FEMALE	12.0	71.0	PARTIC	64.0	NO ABAN	NO ABAN
20110789	MALE	9.0	9.0	STATE	60.0	NO ABAN	NO ABAN
20111426	MALE	13.0	100.0	STATE	65.0	NO ABAN	NO ABAN
20111325	MALE	12.0	76.0	PARROQ	58.0	ABAN	ABAN

ACCURACY	PRECISION	SENSITVITY
0.831858407079646	0.87777777777778	0.9080459770114943

Fig. 5. Classification Performed on Test Set and Performance Metrics.

Considering the values obtained for the established performance measures, it is possible to determine which model is the most appropriate when classifying.

V. RESULTS AND DISCUSSION

As shown in Fig. 6, six different models were implemented. The structure of the model was stored in the database along with the information of its performance metrics, the sizes of the training and test sets; in addition to the variables that were used.

A different model was first created for each algorithm (neural network, ID3 and C4.5) where performance metrics were obtained that were around 80% minimum. For these models, the training and test set of 338 and 113 instances, respectively, was used. These models were aimed at classifying students according to whether or not they abandoned the career.

Fig. 7 shows the decision tree built by the C.45 classifier, which is the algorithm with higher performance measures with respect to accuracy, precision and sensitivity. From this tree model it is possible to determine which credit ratio was the most significant variable, followed by the admission exam score and the average of grades, on the other hand, the type of admission exam is not included within the tree by which is not a significant variable for the model. Preliminary results, related to this approach, have been presented in [13].

Three other models were implemented to classify students according to their type of risk: Incipient Risk, Moderate Risk, High Risk and Very High Risk.

First, the model for the two decision tree algorithms was constructed, neural networks were excluded for this case due to the transformation of the data since the attributes had very varied ranges of values and the different values were not going to be significant.

NAME	MODEL	ACCURACY	PRECISION	SENSITIVITY
NETWORK MODEL	NEURONAL	0.796460176911505	1.0	0.79646017699115
ID3 DESERTION	ID3	0.823008849557221	0.822222222222	0.94871794871794
C45 DESERTION	C4.5	0.831858407079646	0.877777777778	0.90804597701149
ID3 RISK TYPE	ID3	0.585217391304347	0.585217391304	1.0
C45 RISK TYPE	C4.5	0.652173913043478	0.652173913043	1.0
ID3 RISK TYPE	ID3	0.707964601769911	0.707954601769	1.0

Fig. 6. Classification Models Generated.

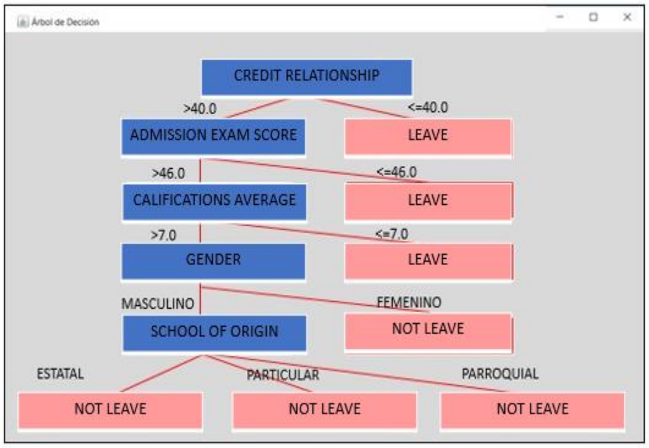


Fig. 7. Decision Tree Generated by the Application to Classify Student Dropout.

To generate these models, only those students who had the "ABANDON" class were filtered from the previous data set, which reduced the test set to 137 and 23, respectively.

The results in the metrics were not very good showing values between 56 and 65%. Analyzing the data, there was a great difference in the distribution of the values of the variables and because of this it did not train the model correctly.

Fig. 8 shows the low percentage of cases for some classes of the variable "DISAPPROVED SUBJECTS" where the greatest distribution is in the class of "NO SUBJECT DISAPPROVED".

A last model was built using the ID3 algorithm and the same input and output variables as in the previous approach with the difference that the dataset was used for the first 3 cases. This time there was a better distribution of training and testing examples, which yields of 0.77, 0.76 and 0.94 were obtained for accuracy, precision and sensitivity, respectively.

Unlike the first approach, in this one the most significant variable is the number of ABANDON SUBJECTS followed by the DISAPPROVED SUBJECTS.

Fig. 9 shows the decision tree generated by the application to classify students according to the types of risk previously defined.

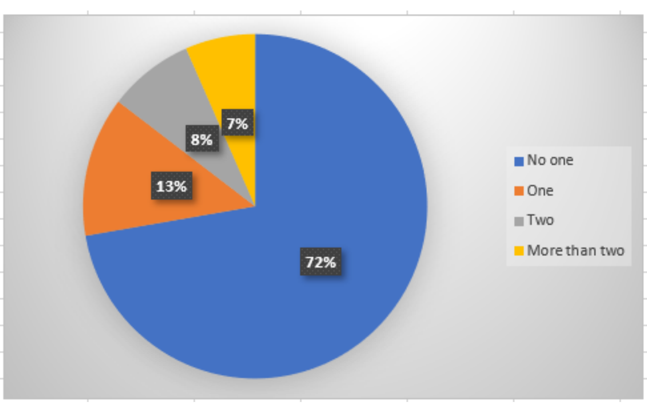


Fig. 8. Distribution of Classes of the Disapproved Subjects (Training Set: 113 Registers).

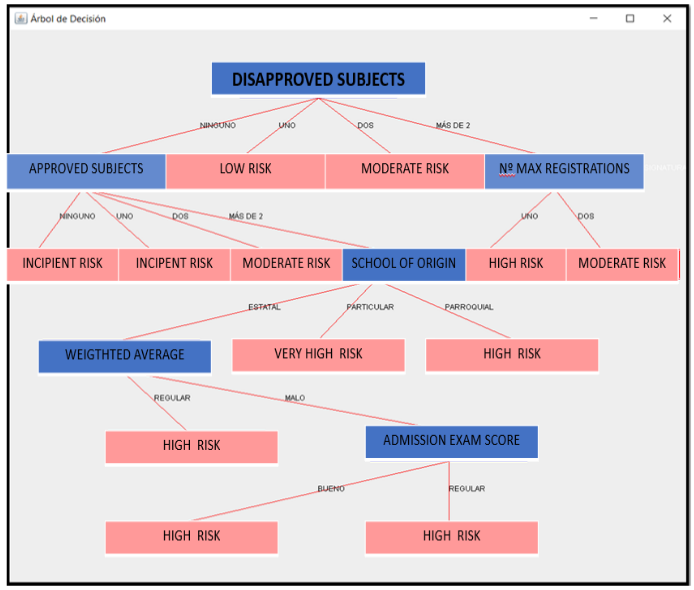


Fig. 9. Generated Decision Tree to Classify Risk Type.

Data Mining Techniques have been shown to be effective tools for obtaining models to predict the permanence of students enrolled in an engineering career, as well as to more specifically determine the risk of dropping out of the university.

VI. Conclusions

With the development of this work, it has been possible to determine those factors that affect academic performance, for this purpose different variables of the personal and academic type were used for two different approaches.

The first approach was used to determine student desertion and three classification models were created (one per algorithm), for which the C4.5 algorithm presented performance improvements with respect to the neural network and the ID3 algorithm (because ID3 did not you can work with continuous data).

A student's current credit ratio with respect to the credits they should possess turned out to be the most significant variable in the construction of the model, followed by the notes, while the type or modality of admission exam with which the student he entered college did not turn out to be significant as he did not appear in the decision tree generated.

The second approach was worked with two different types of training and testing sets. However, the same entry variables were worked on (number of abandoned subjects, number of subjects disapproved, weighted average, admissions test score, home school and maximum number of failed enrolments in a subject). It was also worked with the same output variables (incipient, moderate, high and very high risk).

This second approach achieved lower metrics than using the first approach. A misdistribution was determined in the different values of the student attributes for each set so the experiment was re-performed using the initial dataset which had a better distribution in the data and metrics of acceptable performance to be able to classify the type of risk. The generated tree determined that the most significant variable was the number of abandoned subjects. The performance measures obtained could be improved, as well as the rate of success of the classification of the models by increasing new variables in the future, such as social, economic variables, etc.

VII. Future Work

To expand the results of the research it is proposed to consider institutional and socio-economic variables. At the same time, increasing the number of variables should consider working with a much larger data set, taking information from students from more professional schools at the university.

The elaborate system can be adapted to create models with different input and output variables (classes). This was demonstrated as different variables were used for each approach for both input and output variables (two variables for the first approach and four variables for the second).

The best performing models can be used in several professional schools or each school uses a different classification model that best suits its needs.

Acknowledgment

This work was carried out with the support of our house of studies, the National University of San Agustin, in which the Vice-Chancellor of Research channels the resources from the mining canon and convenes a set of insolvable financial schemes.

It is through one that the IBA 004-2016 project was funded, "Model of Academic Performance Assessment for the Detection of Outstanding Students and Students at Academic Risk".

References

[1] I. Mushtaq and S. Nawaz. “Factors Affecting Students’ Academic Performance”. Mohammad Ali Jinnah University Islamabad, Pakistan. 2012.

[2] S. Oloruntoba and J. Akinode. “Student Academic Performance Prediction using Support Vector Machine”. Computer Science Department, the Federal Polytechnic,ILARO,Ogun State, Nigeria. 2017.

[3] S. Sumam and J. Puthiyidam, “Evaluating Students Performance by Artificial Neural Network using WEKA”, International Journal of Computer Applications (0975 - 8887), June 2015.

[4] I. Ganiyu. “Data Mining: A Prediction for Academic Performance Inprovement of Science Student using Classification”. 2016.

[5] D. Kabakchieva, “Student Performance Prediction by Using Data Mining Classification Algorithms”, University of National and World Economy Sofia, Bulgaria, November 2012.

[6] S. Anupama Kumar and M. Vijayalakshmi, “Efficiency of Decision Trees in Predicting Student’s Academic Performance”, India 2011.

[7] S. Sharma, J. Agrawal, S. Sharma, “Classification Through Machine Learning Technique: C4.5 Algorithm based on Various Entropies”, International Journal of Computer Applications 82(16):28-32, October 2013. DOI: 10.5120/14249-2444.

[8] B. Baradwaj and S. Pal, “Mining Educational Data to Analize Students’ Performance”. International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6. 2011.

[9] C. Menacho, “Predicción del rendimiento académico aplicando técnicas de minería de datos”, Universidad Nacional Agraria La Molina, Lima – Perú, 2017.

[10] N. Zacharis, “Predicting Student Academic Performance in Blended Learning Using Artificial Neural Networks”. International Journal of Artificial Intelligence and Applications (IJAIA), Vol. 7, No. 5. 2016.

[11] K. Manchandia and N. Khare, “Implementation of Student Performance Evaluation through Supervised Learning Using Neural Network”. 2017.

[12] R. Bhardwaj and S. Vatta, “Implementation of ID3 Algorithm”. CSE, Bahra University, India, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.3, June 2013.

[13] N. Bedregal-Alpaca, V. Cornejo-Aparicio and D. Aruquipa-Velazco, “Analysis of the academic performance of Systems Engineering students, desertion possibilities and proposals for retention”, INFONOR 2019 : 10th International Conference on Computing and Informatics in Northern Chile, 2019. In press.

Predicting the Future Transaction from Large and Imbalanced Banking Dataset

Sadaf Ilyas[1], Sultan Zia[2]
Zaib un Nisa[5]
University of Lahore
Gujrat Campus, Gujrat, Pakistan

Umair Muneer Butt[3]
Universiti Sains Malaysia
University of Lahore
Gujrat Pakistan

Sukumar Letchmunan[4]
Universiti Sains Malaysia
Penang, Malaysia

***Abstract*—Machine learning (ML) algorithms are being adopted rapidly for a range of applications in the finance industry. In this paper, we used a structured dataset of Santander bank, which is published on a data science and machine learning competition site (kaggle.com) to predict whether a customer would make a transaction or not? The dataset consists of two classes, and it is imbalanced. To handle imbalance as well as to achieve the goal of prediction with the least log loss, we used a variety of methods and algorithms. The provided dataset is partitioned into two sets of 200,000 entries each for training and testing. 50% of data is kept hidden on their server for evaluation of the submission. A detailed exploratory data analysis (EDA) of datasets is performed to check the distributions of values. Correlation between features and importance of characteristics is calculated. To calculate the feature importance, random forest and decision trees are used. Furthermore, principal component analysis and linear discriminant analysis are used for dimensionality reduction. We have used 9 different algorithms including logistic regression (LR), Random forests (RF), Decision tree (DT), Multilayer perceptron (MLP), Gradient boosting method (GBM), Category boost (CatBoost), Extreme gradient boosting (XGBoost), Adaptive boosting (Adaboost) and Light gradient boosting (LigtGBM) method on the dataset. We proposed LighGBM as a regression problem on the dataset and it outperforms the state-of-the-art algorithms with 85% accuracy. Later, we have used fine-tune hyperparameters for our dataset and implemented them in combination with the LighGBM. This tuning improves performance, and we have achieved 89% accuracy.**

Keywords*—*Machine Learning (ML); banking; Santander; transactions; prediction; imbalanced; unbalanced; skewed; hyperparameter; oversampling; undersampling; EDA; dimensionality reduction; PCA; LDA; LR; RF; DT; MLP; GBM; CatBoost; XGBoost; AdaBoost; LigtGBM

I. Introduction

The invention of ML has become an essential advancement in technology, which has opened several doors to success for all areas. As technology changes rapidly, it is necessary to keep oneself up-to-date in this ever-changing era. The speed of technological advancement has progressed exponentially. Presently, the opportunity has arrived, and machines are getting ready to make judgments at their very own, and it's all because of artificial intelligence or particularly ML [1].

The development of skills guides us to live in an epoch where the achievement is not just about the present but to excel; you have to think like a chess player. The financial sector is one of the substantial areas in terms of value as well as the production of large size of data in seconds. Furthermore, it needs real-time decision making according to the situation [2]. To understand this, we take the simplest example of traditional banking systems, in which one has to visit the bank and give a check to the cashier to withdraw money during the banking hours. In this process, every single transaction was made/processed by humans. But now, with the start of e-banking, the situation has become very complicated. No doubt, e-banking has opened new horizons and helped us in many ways (e.g., ATMs, online money transfer, online shopping, etc.) to make it possible for making transactions whenever we want and wherever we are. This causes an exponential increase in transacted data produced by banking (no time limitation). Over time, e-banking services evolve and increase day by day (Fig. 1) [1].

This new system also causes some problems, e.g., Security, authentication, vulnerability, extensive data, no time limit, monitoring, etc. [3]. To protect the system from these problems seems near to impossible for humans in terms of time, cost, and efficiency. The need for intelligent machines was felt because humans could not achieve the processing of millions of transactions in seconds. To accomplish such goals, ML has been introduced in banking to make decisions on the spot on behalf of humans even when humans are not supervising them. This is not the end. This necessity has become a powerful tool, and much more is being achieved by using it.

This paper consists of techniques to handle class imbalance while improving the accuracy of prediction. Rest of the paper is organized as Related work, Problem statement, Methodology (including data analysis, techniques to handle imbalance and implementation of classification algorithms, dimensionality reduction & resampling), Results, finding Hyperparameters and combined the best-performed algorithm (based on results), Conclusion and Findings & Finally Future Work.

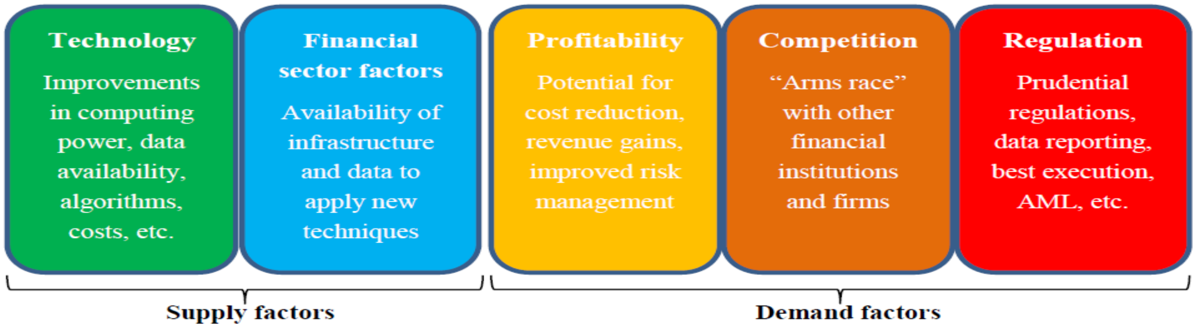


Fig. 1. AI and Machine Learning Factors for Financial Adoption [1].

II. Related Work

Since the ML has been introduced in banking, it has revolutionized this vital sector. The idea of using ML in banking was first given to make financial decisions using neural networks by Hawaly et al. [2], which laid the foundation for the automation of the banking sector. Nowadays, machine learning is not only used for automation, but also for predicting the future, finding customer loyalty, information on standard recovery rates [3], learning optimal coverage rates[4], modeling investor sentiment [5], fraud detection [6], detecting the stock price [7, 8], client maintenance, programmed credit endorsement, extortion discovery, showcasing and hazard the executives in the financial part [9]. Banks keep up a lot of data about their clients. This information can be utilized to make and keep up an unmistakable relationship and association with clients to manage them exclusively for specific items or banking offers. Information mining [10] has turned out to be prevalent for illustrative and prescient applications in financial procedures.

Reverberating its future advantage to banking, various early investigations have additionally shown up in the journal of banking and finance during the 1990s, which investigated the potential for ML to improve loaning choices and credit chance management. Altman et al. [11] connected neural systems to order Italian firms dependent on the probability of budgetary trouble, while Varetto [12] based on this examination by applying hereditary learning calculations to a similar subject. Later research into account, diaries has maintained the attention on the forecast; however, it has moved towards profound learning methods, and other propelled ML procedures. These ongoing applications include: the comprehension of default recuperation rates cheng and cirillo [3] learning ideal choice supporting standards Nian et al. [4] demonstrating financial specialist feeling Renault [5] and the discovery of stock value advancement dependent on request books Kercheval and zhang, 2015 [13]. As fund diaries have made provisional strides to recognize the capability of the new methods of ML in an account, different orders have been attempting progressively strident endeavors to apply ML ways to deal with budgetary information. To some extent, this is because of the appeal of the exhaustive, organized, and effectively open, information accessible in money. Examine on ML and account outside of money diaries are surpassed by an extensive various ML and fund explore in fund diaries.

There can be many factors supporting the use of ML just like other disciplines are using, but being a financial institute, there are also some supply and demand factors. The organization's essential reports are fiscal summaries that dismiss its monetary status doer [14]. The financial proclamation is the fundamental reason for choice-making concerning an immense number of speculators, leasers, and different people needing bookkeeping data, just as a robust articulation of business execution, financial status, and the social obligation of recorded organizations and otc organizations [15].

Be that as it may; lately, instances of deceitful financial proclamations have turned out to be progressively genuine (wells [16]; Spathis et al. [17]; yeh et al. [18]). There have been numerous instances of false financial articulations in us & Taiwan since the Asian financial crisis in 1997. Enron case in 2001, the Abit computer, pro comp, info disc, and summit technology cases in 2004 in Taiwan and WorldCom case in 2003 in us. Given these episodes, it has turned out to be essential to have the option to distinguish fraudulent conduct before its event. Information mining is a critical apparatus for managing complex information investigation and classification. It identifies profitable occasions that are covered up in a lot of information for examination and abridges the information in an organized model to give a reference to essential leadership. Information mining has numerous different capacities, for example, classification, affiliation, grouping, and estimating [19].

Some explorers have been made in the field of client weakening and maintenance investigation in banking segments. A few examinations uncover that the most significant factors impacting client decision are compelling and proficient client administrations, quickness and quality administrations, assortment of administrations offered and little e-administration charges, web-based financial offices, wellbeing of assets and the accessibility of innovation-based service(s), low financing cost on advance, advantageous branch area, picture of the bank, well administration, and generally bank condition [20, 21]. Then again, the client is the center of their activity, so sustaining and holding them are significant for their prosperity. Numerous looks into were hung on client maintenance just as client weakening investigation lift is utilized as an appropriate measure for regular loss examination, supported credulous baye's system, specific baye's system, neural network system, and the above classifiers techniques [22]. Their first spotlights were on constant loss investigation utilizing the lift. It can be determined by taking a gander at the total targets threshold up to p% as a level of everything being equal and isolating by p%. An agitate model with a prophetic presentation and membership set built to help machines [23]. They demonstrated that help vector machines show excellent speculation execution when connected to uproarious showcasing information. The model beats a strategic relapse just when the suitable parameter-determination method is compared, and the arbitrary woodlands outperform SVMs. Soeini et al. [24] chose 300 records of clients and by using clementine software for customer churn prediction. In this paper, statistic factors are used to decide the ideal number of bunches in k-implies grouping and assessed double characterization techniques (quest, c5.0, chaid, cart, Baye's systems, neural network systems) that foresee clients agitate. Goonetilleke et al. [25], in his paper, used DT and NN to create a model that anticipates stir. The presented models are evaluated using the roc curve and AUC deem. They likewise embraced cost touchy learning methodologies to address imbalanced class marks and unequal misclassification costs issues. [26] talked about business bank client stir forecast dependent on the svm model, and utilized an arbitrary testing strategy to improve the svm model, considering the unevenness attributes of client informational collections.

An examination explored determinants of client agitate in the Korean portable media communications administration showcase dependent on client exchange and charging information. Their investigation characterizes variations in a client's position from dynamic to no-use or suspended as partial surrender and from effective use to beat as complete deserting. Results demonstrate that a client's status change clarifies the connection between beat determinants and the likelihood of churn [27]. A Neural Network (NN) based system on dealing with foreseeing client stir in membership of remote cell administrations. Their aftereffects of analyses show that neural network-based methodology can anticipate client stir with precision over 92% [28]. A scholastic database was built between 2000– 2006 covering 24 diaries and proposes a grouping plan to arrange the articles. Nine hundred items were recognized and checked on for their immediate significance in applying information mining strategies to CRM. They found that the examination territory of client maintenance got most research consideration, and order and affiliation models are the two regularly utilized models for information mining in CRM [29]. A study on the idea of data mining and customer relationship management in sorted out banking and retail enterprises was additionally talked about [30].

Banking fraud is a profound term having many types. Zhou et al. [31] have studied several traditional machine learning algorithms, for fraud detection, they choose an improved version of gradient boosting decision tree within bankcard enrollment on the mobile device based payment for use in a real system, namely, XGBoost. K. Seeja and m. Zareapoor [6] proposed an intelligent model that can efficiently detect credit card fraud of data sets of anonymous and highly unbalanced transactions for credit card transactions using frequent extraction of sets of elements.

To find the fraudulent in the incoming transaction of a particular client is closest, an adjustment algorithm is also proposed so that a decision can be made accordingly. Furthermore, they used SVM, NB, KNN& RF on the same data, and when compared, the proposed model performed better than all of these. Another plus point for this model was that it could handle class imbalance too. Cheng et al. [32] proposed a fraud detection framework based on CNN that has intended to record the intrinsic patterns of fraud behavior extracted from tagged data. The data of abundant transactions are represented by a matrix of functions, to which a convolution neural network is applied to identify a series of underlying patterns for each sample. On the other hand, [33] the researcher has conducted research and analyzed fraudulent data by applying DT, BBN, SVM, and ann. In this research, the author has used two stages of statistical treatment to produce more accurate results, unlike other single-stage algorithms.

Many efficient and effective techniques have been presented based on the decision tree for classification [34] and prediction. There are many algorithms to build a decision tree model [35]. The first stage consists of the most influential variable selection by using two primary decision tree techniques, cart & chaid. Whereas, in the second stage DT (cart & chaid), BBN, SVM, and ANN are used to detect fraudulent transactions. Sezer et al. Proposed an image recognition technology to predict technical stock patterns [7]. The proposed model is a CNN model that performs five functions, dataset pre-processing (extract/transform), data labeling, image formation, and CNN performance. The goal of this practice is to determine the best fit for the buy, sell, and hold points in the time series of the associated stock prices. However, [8] the researcher has used online data to make a knowledge base by combining with ensemble methods, e.g., Neural network regression ensemble (NNRE), support vector regression ensemble (SVRE), boosted regression tree (BRT) and random forest regression (RFR) to predict short-term stock prices.

Moro et al. [36] worked with a vast arrangement of information gathered from 2008 to 2013 from a Portuguese retail bank, including the ongoing money related emergency. They broke down many capacities identified with the qualities of the bank's customers, items, and economics. Because of a choice of self-loader works that were researched in the displaying period of their strategy, performed with information before July 2012, the informational collection was diminished to 22 capacities. They likewise thought about four dm models (logistic regression (LR), decision trees (DT), neural network (NN) and support vector machine (SVM)) utilizing two measurements (region of the collector working bend (AUC) and territory of the total bend of lift) from which NNexhibited the best outcomes (AUC = 0.8 and lift = 0.7), with the goal that 79% of the endorsers could become to by choosing the best-positioned customers.

As per turban et al. [37], business intelligence incorporates models, devices, databases, applications, and philosophies to utilize information to help the choices of business chiefs. Information mining is a business insight innovation that utilizations information-driven models to remove helpful learning, that is, examples of intricate and enormous information accumulations [38]. Chitra and subashini [39] have utilized a few information extraction calculations for client maintenance, programmed credit endorsement, misrepresentation location, advertising, and hazard the executives in the financial division. They have distinguished a few strategies and models to improve client maintenance and misrepresentation discovery.

Hu [22] applied information mining strategies to assist retailers with weariness examination to recognize the scope of customers with a high likelihood of acting. He utilized the choice tree (DT), guileless invigorated bayesian system, particular bayesian system, the neural system as an information mining model. Ghosh et al. [40] from a company that implements a specific "bonus program". 2 algorithms have been used (Naive Bayes improved with ADA and c4.5 developed with ADA with CF) that too generates probability. Tool for credit scoring is designed to speed up loaning judgments, while possibly preventing incremental risk with the help of machine learning. In any case, money lenders or banks are increasingly going to extra, amorphous and semi-organized information means, containing online life action, usage of cell phone and instant message action, to get a more nuanced perspective on reliability, and improve the score exactness of credits.

Implementing AI computations to this set of stars of new information has empowered assessment of subjective factors, for example, employment conduct and readiness to pay. The ability to use additional information on such measures takes into consideration a more significant, faster, and inexpensive partition of mortgagor quality and ultimately stimulates a quicker credit choice [41]. Ling and li [9] utilized information-digging procedures for direct promoting in three informational collections from three distinct sources. The primary informational index was for the advancement of an advance item in Canada. The following informational collection was from a considerable disaster protection organization, and the third informational index was 260 s. A great deal of work has been finished utilizing ML in banking, and many have utilized distinctive managed, solo and troupe strategies to accomplish their objectives, be that as it may, some have used to consolidate regulated and unaided procedures, for example, e.g., Wang et al. [42] have proposed a half breed procedure for credit scoring of the clients.

Table I shows the details about the banking-related tasks and the algorithms used to perform these tasks.

TABLE. I. MACHINE LEARNING TECHNIQUES USED IN BANKING TO ACHIEVE DIFFERENT GOALS

Task	Methodology	Year	Reference
Fraud Detection & Prevention	(SVM, KNN, NB, RF) , CNN, CNN, Clustering	2014,2016, 2016, 2013	[6], [32],[33], [39]
Stock Prices	CNN, (NNRE, SVRE, BRT, RFR)	2018,2018	[7], [8]
Risk Management	Bayesian Model	2017	[43]
Crisis Management	LR, DT, NN, SVM	2011	[36]
Customer Attrition	DT	2005	[22]
Banks Retailing		2010, 2005	[37], [38]
Customer Retention	DT	2013	[39]
Credit Approval	(DT, SVM, LR), DT	2015	[39], [44]
Marketing	AB, NB	1998	[9]

III. PROBLEM STATEMENT

A US-based bank named Santander has recently conducted a competition with the help of an ML and data science competition site KAGGLE. The bank shared its data by keeping the attributes hidden. This is a binary classification competition; in this competition, we need to predict whether a customer will do the transaction or not. The dataset that is shared by the organizer has numeric data fields.

IV. WORK / METHODOLOGY

In this research we have worked on a recent dataset of a bank, analyzed it for distributions of mean, standard deviation, skewness, and kurtosis, check for feature correlations and importance and tried to reduce dimensions of data by implementing principal component analysis (PCA) as well as linear discriminant analysis (LDA). Furthermore, to tackle class imbalance, a variety of methods, including evaluation metrics, a variety of different classifiers, and resampling, are used. Moreover, to increase the overall prediction, hyper-parameters are also computed for the best performing algorithm in combination with the most suitable evaluation metric for our dataset.

All the implementations are done using python 3.7.1 on anaconda3 & Jupiter notebook 5.7.4. Also, the machine used in this whole process is a Fujitsu core i3 2nd generation. All the algorithms are taken from sci-kit learn [45].

A. Dataset

The dataset [46] was taken from the Santander transaction prediction competition from the KAGGLE site. Dataset set has two different files, consisting of 200,000 entries each. The difference in both files was that one file has the target too, to use for training the model while the other file didn't have a target variable that can be used for making predictions. To check the prediction results, one can upload the results on the competition site to see how accurate the predictions are made. Both the datasets had an id column as well as 200 variables to make predictions while the training dataset one extra column consisting of the target. According to the information, the dataset has 800,000 entries in total, which is quite large [47].

B. Pre-Processing

Data pre-processing is a process that manipulates raw data for further implementations on the dataset. It involves handling incompleteness, inconstancy, and errors in data. In real-world data are usually imperfect, varying, duplicate, and noisy. It may have lacking attribute values or may contain aggregate data only. There can be errors or outliers in data that may change affect the overall distribution of data, repeated values, and differences in attribute names can also be problematic. As these issues of data may lead us to wrong predictions for this, we use data pre-processing because it is an excellent approach to resolve such matters before occurring. It was seen while pre-processing that the dataset is entirely clean, with not a single missing value in both train & test datasets. No inconsistency, redundancy, or noise found. Furthermore, all the attributes have floating-point (numerical) values, and they're not a single categorical attribute.

C. Exploratory Data Analysis (EDA)

Exploratory data analysis, well-known as EDA, is a tactic for examining datasets to summarize their core features, often with realistic approaches. EDA is used for getting summarized insights about data to provide us in-depth knowledge about is which in the future we can use in the modeling task. These insights can be graphical as well as non-graphical. But graphics give us more details as compared to non-graphic approaches; graphical representations of data are preferred more. It is tedious and time-consuming to look at a value of attributes or a whole spreadsheet and conclude the vital characteristics of the data. It might also be annoying and overwhelming to extract insights by seeing ugly numbers. EDA techniques have been developed as an aid in this situation. Chris Chatfield [48] strongly encouraged to use of EDA to describe the data and formulate the model based on this data.

In this phase, we started by checking the min and max of data. Min and max refer to the minimum value as well as the maximum value. It was discovered that the datasets have the data types far more than the requirements of the data, hence consuming more memory not just on disk, but it requires more time to process the data (Fig. 2 and Fig. 3). To handle this data more effectively and to reduce the memory and time cost, data types have been changed with the help of memory reduction function.

This worked well and helped in saving almost 50% of the memory occupied by both datasets, as this process repeated once again but with the other dataset, to be used for testing. Reducing the memory cost will also reduce the computation power & time required. Table II and Table III are given below showcase the dataset along with the memory before and after memory reduction of both train and test datasets.

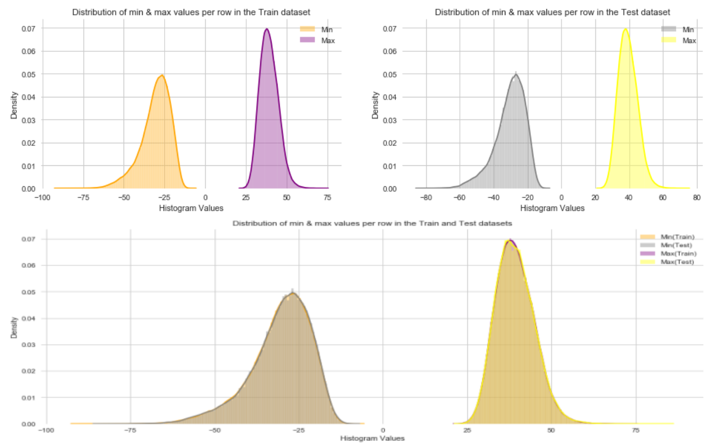


Fig. 2. Distribution of Min and Max Values Per Row in Train & Test Datasets.

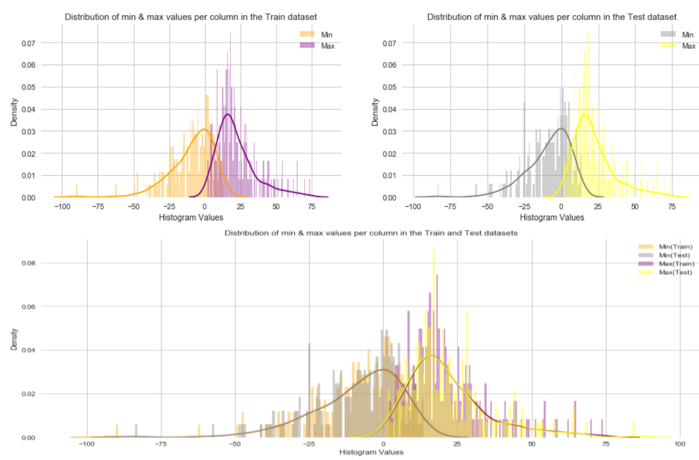


Fig. 3. Distribution of Min and max Values Per Column in Train and Test Datasets.

TABLE. II. REDUCING MEMORY OF TRAINING DATASET BY CHANGING DATA TYPES

Training Dataset				
Columns	Before Preprocessing		After Preprocessing	
	Data types	Memory Utilized	Data types	Memory Utilized
ID_code (1)	Object	308.2276153564453 MB	Object	154.30458068847656 MB
Variables (200)	Float64		Float32	
Target (1)	Int64		Unit8	

TABLE. III. REDUCING MEMORY OF TESTING DATASET BY CHANGING DATA TYPES

Testing Dataset				
Columns	Before Preprocessing		After Preprocessing	
	Data types	Memory Utilized	Data types	Memory Utilized
ID_code (1)	Object	306.7017364501953 MB	Object	154.1138458251953 MB
Variables (200)	Float64		Float32	

Furthermore, a complete analysis of datasets was conducted in which datasets have been analyzed for four major components of descriptive statistics; mean, standard deviation (sd), skewness and kurtosis, and their distribution in the train as well as test dataset, which showed that both datasets are almost identical. But an important thing finds out during analysis was that the distribution of class 0 and 1 is skewed, making this data imbalanced.

D. Feature Correlation

Feature correlation is another statistical method that can check whether the features are related to each other and, if so, how strongly these are associated. The name "correlation" denotes a universal relation among variables. Frequently, correlation is considered as the paramount step to understand these relations for building better statistical models. It can provide support in foreseeing one variable from another. It shows the existence of a contributory association, which is also used as a critical point and a base for several other modeling methods.

During analyzing the correlation between variables, we find out that there is not much correlation among these, neither positive nor negative, and the variables of both datasets are independent. Variables that are correlated, either positive or negative, are known as the dependent variable. Figures depict the correlation between the variables of each train & test datasets that how many variables are correlated. The scale on the right shows the correlation level and color. As there is not too much correlation or least correlation, that is why the color of the variables are in black. The only light color in figures is the correlation of a variable with itself (Fig. 4 and Fig. 5).

E. Feature Importance

Datasets were also analyzed to get knowledge about the essential features of these, which participates more in the decision-making process. For these two different techniques were used to see the vital features of data as well as the working of classification methods on the same data. Decision tree and random forest classifiers were selected for this task. By finding the importance of features, it was found out that “var_110“ is an essential feature when working with a random forest classifier while the “var_86“ has the least importance. On the other hand, when working with a decision tree to find out essential features, a tree was formed, showing 6 levels starting from root to leaf nodes. The decision tree finds “var_81” is the most important one hence build the complete tree by selecting it as the root node.

F. Dimensionality Reduction

Variables or features of data are also known as dimensionality, and dimensionality reduction refers to dropping the no. Of variables and finding the least no. Of variables to analyze data more effectively because the complexity increases with the increase of dimensions. In ML, the problems caused by working with a higher number of aspects are known as the curse of dimensionality. So to diminish the complication and make it easier to work on a dataset having a large no. Of the dimensions, different methods are used. In this regard, a well-known technique is principal component analysis (PCA) that lets us review and to envision the facts in a dataset holding interpretations defined by multiple intercorrelated measurable variables by defining new variables or dimensions called principal components.

- Principal component analysis (PCA): The foremost objective of PCA is to recognize guidelines along which the deviation in the data is most. It is valuable when we have got data on a large number of features and trust that there is some redundancy among these features. Redundancy, in this case, refers to that more than one variable is correlated might measure the same construct. Based on this, we consider that we can shrink the observed features into a relatively small number of principal components, which will participate for maximum deviation in the observed features [49]. These newly created variables relate to a linear mixture of the original variables. These variables will always smaller in numbers as compared to the originals.

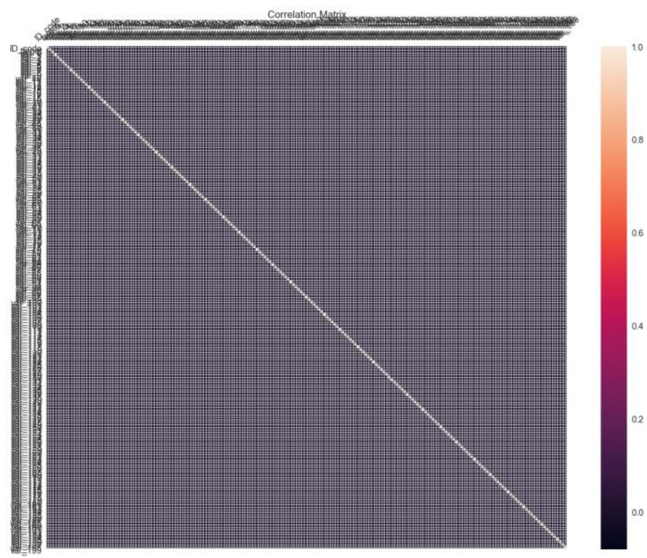


Fig. 4. Correlation between Features of Train Dataset.

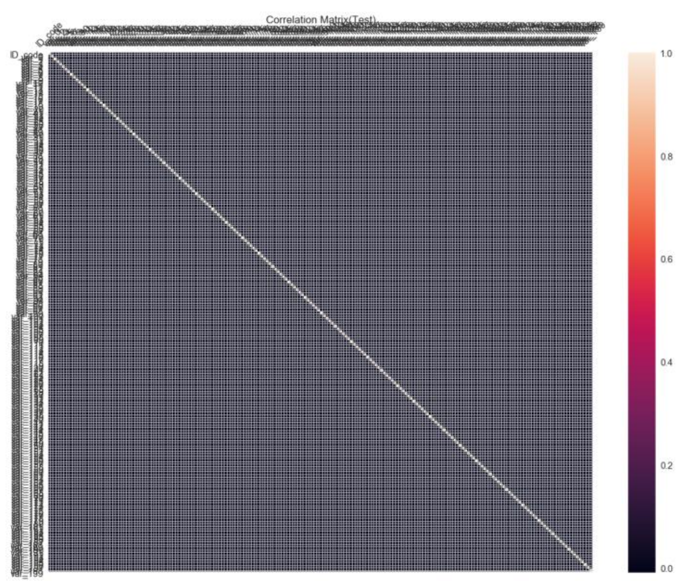


Fig. 5. Correlation between Features of the Test Dataset.

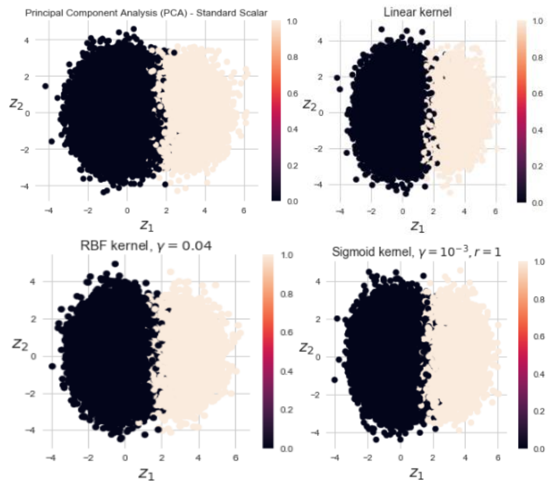


Fig. 6. Finding the Principal Component by using 4 different Techniques.

As can be seen, none of the applied PCA techniques could separate the "class0" & "class1" properly, not even reducing the dimensions to the least. Hence our PCA failed to meet the goal help us reducing the proportion to assist us in finding the principal components (Fig. 6).

- Linear discriminant analysis (LDA): It is worth mentioning here that after failing in achieving optimum goals for PCA, the researcher has tried another method to reduce the dimensionality, which is linear discriminant analysis (LDA) [50]. Both LDA and PCA are direct transformation techniques, but PCA is an unsupervised technique contrary to LDA, supervised. The LDA is quite similar to PCA. But as PCA helps in getting the axes of components that maximize the deviation of data, LDA helps in finding the axes that maximize the parting among classes [51]. LDA makes predictions assuming Gaussian distribution by taking the mean value for every type and considers variants [52]. The goal is to map a dataset with small dimension space but with good class separation to avoid overfitting along with reducing the computational cost. The plus point of LDA has over PCA is that it tackles overfitting too.

Unfortunately, just like PCA, LDA also couldn't help us in feature subspace, which could maximize the separability of our classes. During analyzing the data, another thing was noticed that this is a binary problem as the target may have only two conditions, either the client will make the transaction (1) or not (0), but there was an imbalance (Fig. 7).

G. *Class Imbalance*

It was found while analyzing data that this dataset has only 2 classes 0 & 1, making it as a binary problem the dataset is imbalanced. As both classes are not proportionate, there are only two classes, and both classes were not properly balanced, as out of 200,000 entries, 179,902 (89.95 %) had the target as 0, while the rest of the entries had the target value 1 which hardly make 10.05 % of aggregate data. The graph below shows the distribution of the training dataset. Class imbalance, as its name depicts, is the unbalanced distribution of classes [53] [54] (Fig. 8).

To address this problem and to get the maximum performance out of any algorithm when predicting, a variety of techniques are used:

- Collecting more data: The major and leading step in avoiding class imbalance is to collect as much data as possible. But in our scenario, the data is already quite enough, as well as is the property of a bank and cannot be collected directly due to several reasons, e.g. Lack of resources, lack of access to clients and lack of information about the parameters or variables.
- Evaluation metrics: Another way to avoid this problem is to consider the performance measure, which can help us not to prevent biases and predicting and understanding the actual performance of classifiers [54]. These performance measures include area under the curve (AUC) and confusion matrix etc. As the class "0" have 90% of the population, if a classifier does nothing and says that 100% data belongs to class "0", we still get 90% accuracy. So, accuracy is not a good measure while working with class imbalance. On the other hand, the confusion matrix shows the results as:

True positives (TP)

Identified as positive that is positive

False positives (FP)

Identified as positive that is negative

True negatives (TN)

Identified as negative that is negative

False negatives (FN)

Identified as negative that is positive

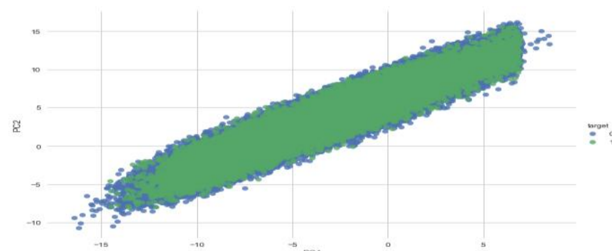


Fig. 7. Linear Discriminant Analysis for Santander Dataset.

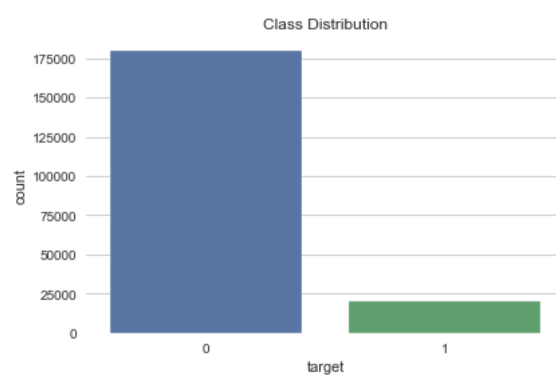


Fig. 8. Bar Graph Representation of the Class Distribution.

Based on these outcomes, we can check the effectiveness of our model by using recall, precision & f1 score. Recall is the number of true positives (TP) divided by the number of true positives (TP) plus the number of false negatives (FN). Recall is also known as the true positive rate (TPR) & sensitivity. Whereas, precision is defined as the number of true positives (TP) divided by the number of true positives (TP) plus the number of false positives (FP). Precision is also called true negative rate (TNR) & specificity (Eq. 1 & 2).

$$Recall = \frac{TP}{TP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

There is a trade-off in the metrics when we increase the recall, precision decreases. To find an optimum combination of precision and recall, we can combine the two parameters using what is called the f1 score (Eq. 3).

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{3}$$

The outcomes of the binary classification can be represented visually through confusion matrix (cm), receiver operating characteristics (roc) & area under the curve (AUC). Cm depicts the actual and predicted labels from a classification problem as a matrix, having actual classification on one side and predicted on the other (Table IV).

On the other hand, roc & AUC are graphical representations of true positive rate (TPR) & false-positive rate (FPR). Roc curve plots the true positive rate (TPR) versus the false positive rate (FPR) as a function of the model's threshold for classifying a positive and AUC metric is used to calculate the general performance of a classification model based on the area under the roc curve (Fig. 9).

- Classification algorithms: Classification methods are a vital part of data mining & machine learning applications. We chose different classifiers because of their different approaches to solve a problem. Every classifier cannot perform best on every type of dataset. A classifier outperforming others on a dataset is not meant that this classifier can outperform the others on any other datasets too.

So, we took some well-known classifiers (including simple as well as ensemble methods) to classify whether a customer will make a transaction or not. These classifiers include logistic regression (LR), random forests (RF), decision tree (DT), multilayer perceptron (MLP), gradient boosting method (GBM), category boost (CatBoost), extreme gradient boosting (XGBoost), AdaBoost and light gradient boosting process (LighGBM). Some of these gave a relatively good performance, while others hardly crossed the 50% accuracy. Among all these, LighGBM outperformed every other classifier. Initially, we did not specify any parameter for classification. We applied the most straightforward implementation to check the behavior of classifiers on our dataset. The roc curves of the algorithms mentioned above are shown (Figs 10-18).

When applied LighGBM, initially, it responded as a regressor and gave floating-point values instead of clear cut 0' or 1's. It is because when we imported LighGBM, we used it as it is, just like other algorithms. But the AUC was remarkably higher than all the previous classifiers applied to our dataset. Also, as the predictions did not classify the data in 0's or 1's, instead, it gave the floating-point predictions like regression; other metrics could not be calculated.

TABLE. IV. CONFUSION MATRIX

Confusion matrix		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

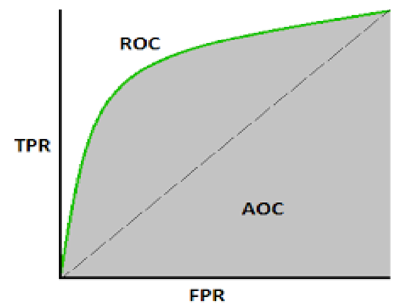


Fig. 9. The Area under the Curve (AUC).

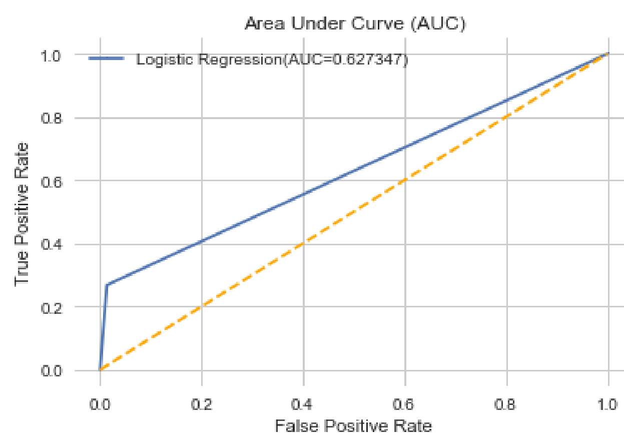


Fig. 10. AUC for Logistic Regression.

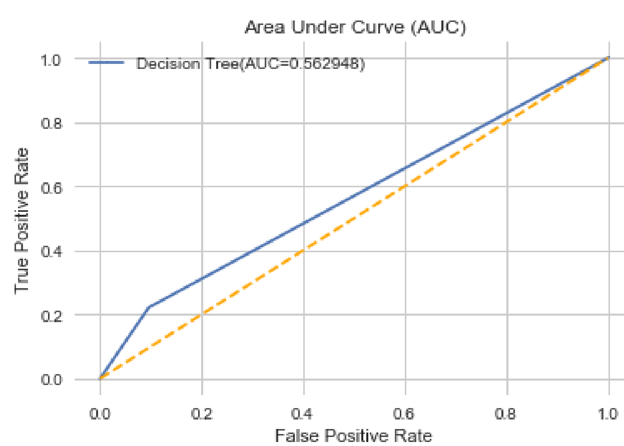


Fig. 11. AUC for Decision Tree.

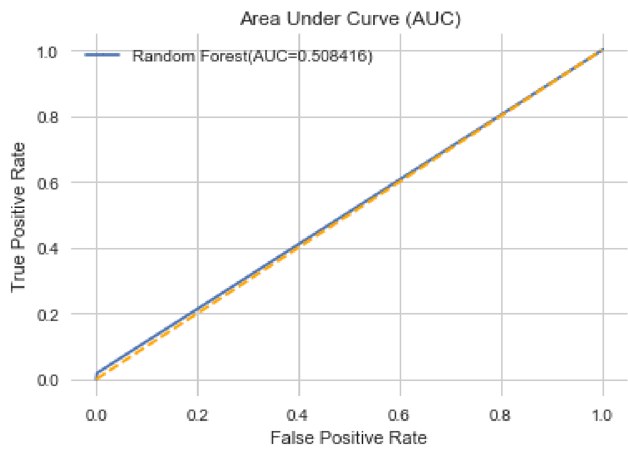


Fig. 12. AUC for Random Forest.

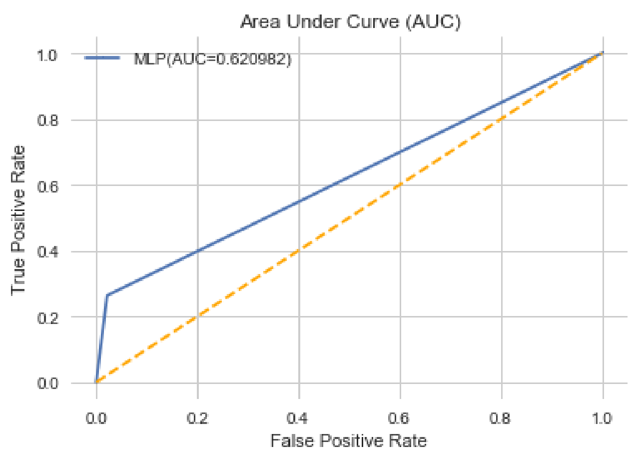


Fig. 13. AUC for Multi-Layered Perceptron.

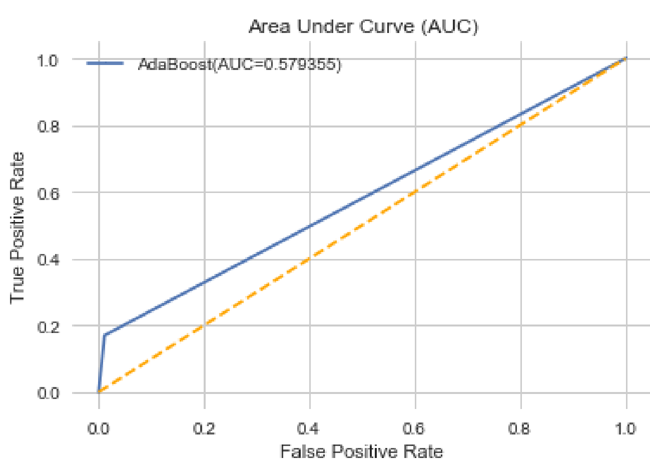


Fig. 14. AUC for AdaBoost.

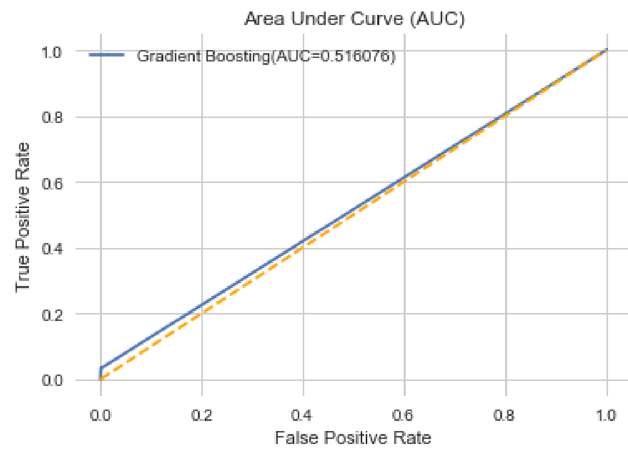


Fig. 15. AUC for Gradient Boosting Method.

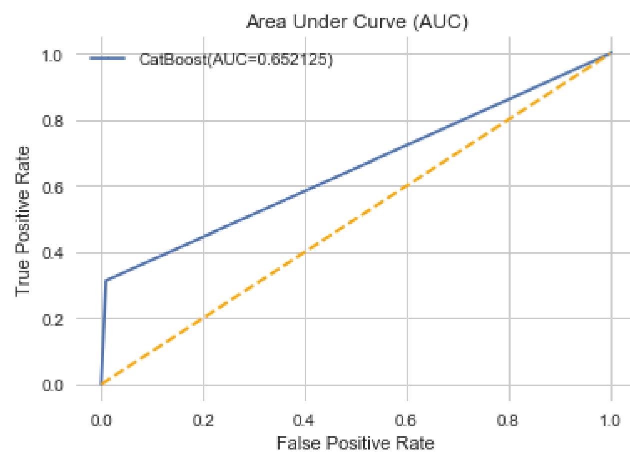


Fig. 16. AUC for CatBoost.

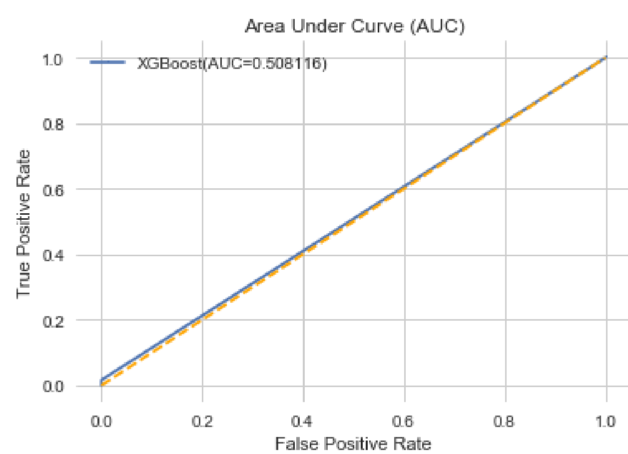


Fig. 17. AUC for XGBoost.

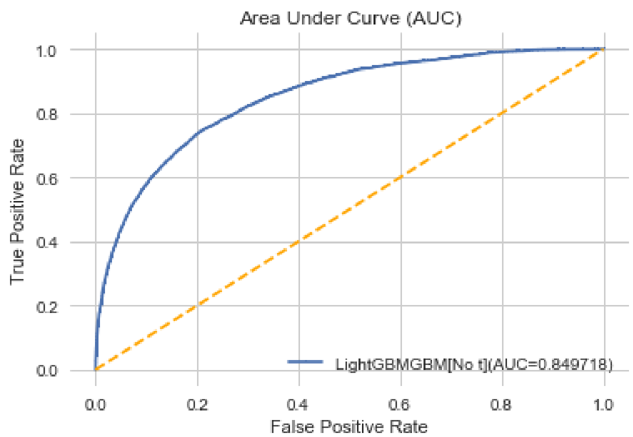


Fig. 18. AUC for LighGBM.

To calculate the values for other metrics and classify data in 0's & 1's, the researcher finds out the threshold after thorough analysis as well as researching and finds out the optimum value for threshold, which was round about 0.1025. By putting this value as threshold converted the predictions to classes 0's and 1's and out the precision, accuracy, recall, and f1 score, but the AUC dropped to 0.748172 (Fig. 19).

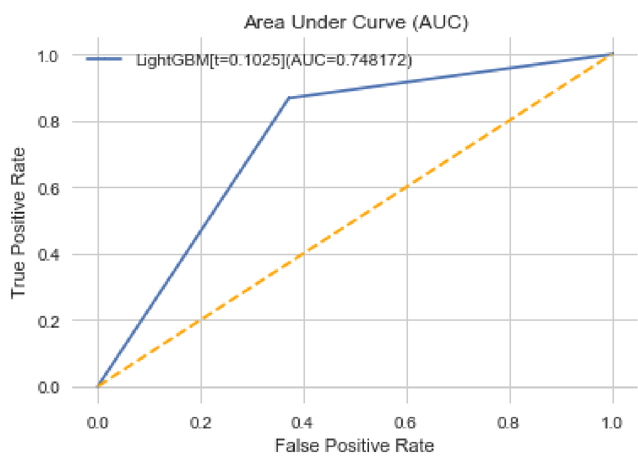


Fig. 19. AUC for LighGBM (Threshold=0.1025).

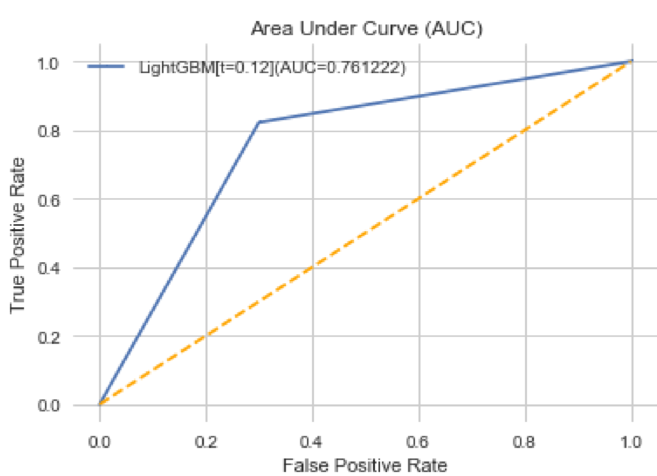


Fig. 20. AUC for LighGBM (Threshold=0.1200).

Further improvement in threshold value ended up on 0.1200 and gave somehow relatively good AUC (0.761222) but still less than without the threshold value (Fig. 20).

To get the binary output, we tried the LighGBM classification method too, which didn't perform well (Fig. 21).

One important thing we noticed here was that the problem was a binary problem; we had to predict the class of data as 0 or 1. But when we uploaded this file to the competition site, it still entertained the submission file, and even we got 85% AUC when submitted. So, as our ultimate goal was to maximize AUC, we kept using LighGBM as regressor because out of all these algorithms tried, LighGBM produced the best results, and the difference was remarkably high. The reason behind this is that LighGBM has a different structure, which leads it outperforming every other classifier tried on this dataset, including the other boosting methods too.

- Resampling: Resampling is a process of removing the imbalance of the classes either by increasing the number of smaller classes equal to the larger class or by decreasing the number of observations from larger classes to be equal to the smaller class. These are called undersampling & oversampling.
    - In undersampling, some of the observations from the majority class are deleted randomly to match the count with the minority class (Fig. 22).
    - In Fig. 23 results of LighGBM on the undersampled dataset can be seen.

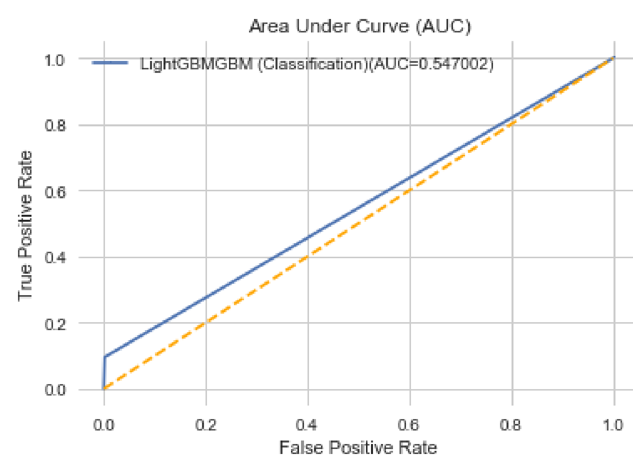


Fig. 21. AUC for LighGBM (Classification).

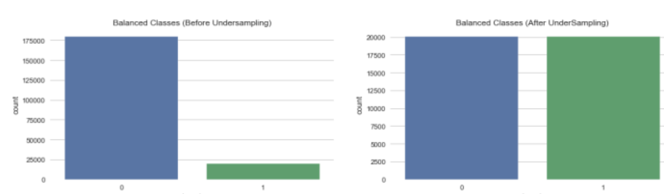


Fig. 22. Comparison of Imbalanced and Balanced (undersampled).

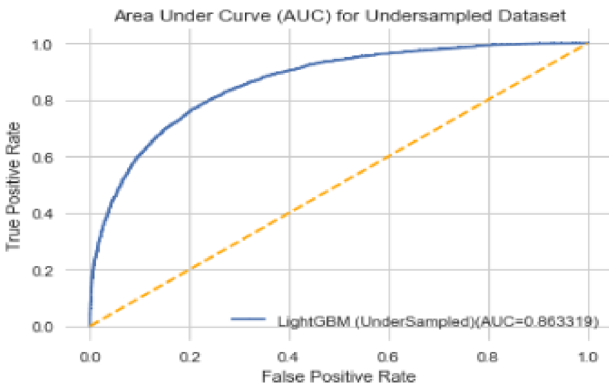


Fig. 23. LighGBM on the undersampled Dataset.

- Oversampling: On the other hand, oversampling is a process that is a little more complicated than undersampling. It is the process of increasing the observations of minority class to a level that matches the majority class (Fig. 24). This can be done by randomly copying the existing observations or generating artificial data that arbitrarily make a sample of the attributes from observations in the minority class. The frequently used technique is called synthetic minority over-sampling method or smote [55]. In simple terms, it looks at the feature space for the smaller class data points and considers its k nearest neighbors.
    - In Fig. 25 results of LighGBM on the undersampled dataset can be seen.

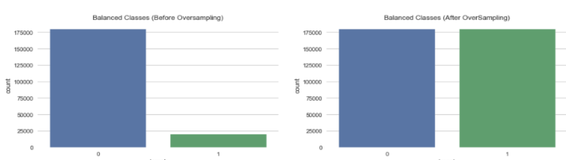


Fig. 24. Comparison of Imbalanced and Balanced (Oversampled).

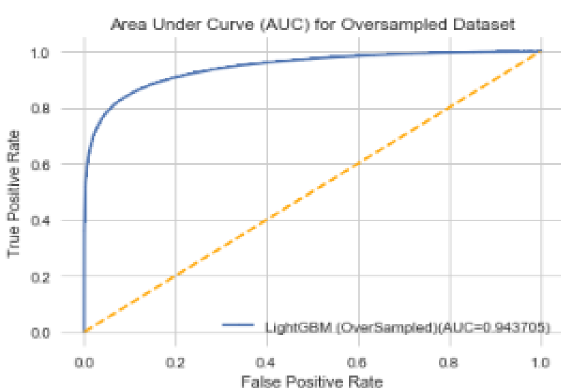


Fig. 25. LighGBM on the Oversampled Dataset.

V. RESULTS

When initially implemented the classifiers, we got the results, as shown in Table II. Among all the nine classifiers, some performed better than others while some hardly crossed the random guessing limit or 50% accuracy. When we submitted the results of these predictions to the website for cross-checking, we got almost the same results there. Among all these random forests performed the least giving "0.50006" while LighGBM outperformed all classifiers tried in this research. LighGBM produced floating-point values as a target or solved it as a regression problem, which gave an excellent AUC (85.0000 %), but we could not find the training score, validation score, precision, recall & f1 score. For this, we went to set the threshold value to convert the target values to either 0 or 1. After thoroughly analyzing the predictions, we got that the threshold is somewhere near to "0.1". So we tried some values and observed the difference we got some excellent results (0. 748172 %) when we set the threshold (t=0.1025) while (76.1222 %) setting the value t=0.1200. These results of LighGBM were still better than every other classifier but significantly less than the results of LighGBM without setting the threshold value.

As we know that the AUC is the trade-off between the precision and recall, one increases when the other decreases. The results of precision and recall show that the implementations other than LighGBM show the precision values higher than recall except for decision tree, which had both the precision and recalls almost the same. But contrary to this, when implemented the LighGBM, we got the high recall values than precision. Getting that LighGBM is working well on our dataset, we selected this for further implementation on resampled datasets. When created resampled datasets, we first implemented the LighGBM once again now on the undersampled dataset. We got an increase in AUC by getting the (86.3319 %). The results were quite good, but we decided to implement LighGBM on an oversampled dataset too, which further gave us a 4.47% increase in AUC, and our AUC reached an excellent 94.3705 %. But when uploaded this result to the website, the result was slightly more than 73% (73.039%). This is because the dataset was synthetically oversampled by smote, which led our model to overfit it (Fig. 26 and Table V).

VI. HYPER-PARAMETERS

To further improve the accuracy, we decided to select parameters, and instead of randomly guessing and manually adding the values, we decided to find hyperparameters for LighGBM. We have chosen these parameters and a range of benefits to test. This code was supposed to run for 1000 times, but it is about testing permutations and combinations on a large dataset, which is quite hectic and time-consuming. It took a long time to just run for 56 times out of 1000, giving about 89% AUC (Table VI).

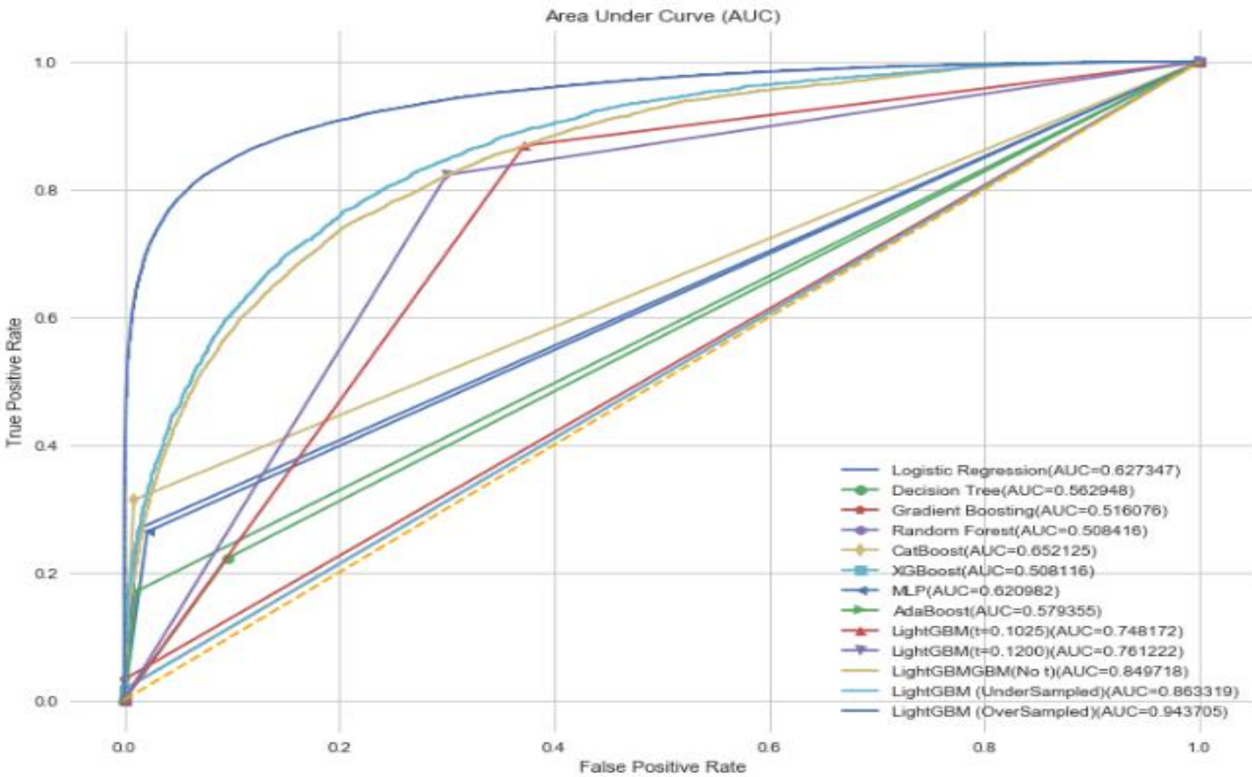


Fig. 26. Comparison Oversampled and undersampled Implementation with all Techniques.

TABLE. V. RESULTS OF CLASSIFICATION ON SANTANDER DATASET

Results of implementation of ML algorithms on Santander dataset (train=200,000, test 200,000)								Submission results
Name	**Training**	**Validation**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Roc AUC**	
Logistic regression (LR)	0.913927	0.913573	0.915600	0.687597	0.268056	0.385735	**0.627347**	**0.63159**
Decision tree (DT)	1.000000	0.832773	0.837200	0.202383	0.219907	0.210781	**0.562414**	**0.56514**
Gradient boosting(GBM)	0.903620	0.902160	0.903820	0.852632	0.032774	0.063121	**0.516076**	**0.51807**
Extreme gradient boosting (xgb)	0.901627	0.900760	0.902620	0.920455	0.016387	0.032200	**0.508116**	**0.50933**
Category boost (CatBoost)	0.935487	0.921140	0.923900	0.790306	0.313372	0.448790	**0.652125**	**0.65006**
Multi layers perceptron(MLP)	0.927993	0.892720	0.906380	0.565174	0.229820	0.326765	**0.605211**	**0.61171**
Random forest (RF)	0.985287	0.899267	0.901140	0.500000	0.014971	0.029071	**0.506664**	**0.50006**
Adaboost	0.907300	0.905893	0.907660	0.620206	0.170140	0.267027	**0.579355**	**0.57989**
LighGBM (classifier)	0.916753	0.907507	0.909080	0.861566	0.095691	0.172251	**0.547002**	**0.55095**
LighGBM as regressor								
LighGBM Without threshold (no t)	-	-	-	-	-	-	**0.850000**	**0.85201**
LighGBM Threshold(t): 0.1025	-	-	0.651960	0.203935	0.868096	0.330280	**0.748172**	**0.85201**
LighGBM Threshold(t): 0.1200	-	-	0.712160	0.231242	0.822375	0.360980	**0.761222**	**0.85201**
LighGBM implementation of resampled data								
Under-sampled dataset	-	-	-	-	-	-	**0.863319**	**0.86417**
Over-sampled Dataset	-	-	-	-	-	-	**0.943705**	**0.73039**

TABLE. VI. RESULTS OF LIGHGBM WITH HYPERPARAMETERS ON SANTANDER DATASET

Results of Implementation of ML Algorithms on Santander Dataset (Train=200,000, Test 200,000)								Submission Results
Name	**Training**	**Validation**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC AUC**	
LighGBM (with HP)	-	-	-	-	-	-	**0.887997**	**0.88856**

Here is the set of best values for these parameters.

Space (best): ['bagging_freq': 3, 'bagging_seed': 100000, 'boost_from_average': 'false', 'boosting_type': 'dart', 'class_weight': none, 'colsample_bytree': 0.4, 'learning_rate': 0.45, 'max_bins': 60000, 'max_depth': 1, 'metric': 'auc', 'min_child_samples': 35, 'min_data_in_leaf': 8, 'min_sum_heassian_in_leaf': 4, 'n_estimators': 209.0, 'num_iteration': 280000, 'num_leaves': 45, 'objective': 'binary', 'reg_alpha': 4.4094144078689945, 'reg_lambda': 1.0182413699039161, 'seed': 100000, 'subsample_for_bin': 340000, 'tree_learner': 'serial', 'verbosity': 1].

VII. CONCLUSION AND FINDINGS

Working on the imbalanced dataset and by implementing 9 different basic, advanced and ensemble classification algorithms on the Santander customer transaction prediction dataset provided by the Kaggle, we find out that selecting the metrics is the first and foremost step to know what exactly we want to get from the classifier when working on Imbalanced data. After that, we can select different classifiers, but we find out that LighGBM performed better on this particular dataset. While working with a large dataset, undersampling can perform well, but the AUC will remain less than as compared with the AUC of hyperparameters find from the original dataset; on the other hand, oversampling or, more specifically, smote may lead to overfitting. Furthermore, the performance of a classifier can be increased by finding the hyperparameters. Finding the best parameters randomly and manually is a long, sturdy, and impossible task. Even by using libraries to find the best parameter values is a hectic and time-consuming job. There can be too many combinations of values to try as well as the size of the dataset matters a lot. When finding hyperparameters, undersampling can be considered, but oversampling, in this case, is not recommended due to a lot of time as well as computation requirement. Moreover, the problem was described as binary (classification) but could also be solved by regression as regression results were also accepted.

VIII. FUTURE WORK

In the future, we’ll implement LightGBM on some other structured imbalanced datasets to find out the scalability. Furthermore, to find more specific and best hyperparameter values, we’ll re-run this code on the latest machine to improve performance.

REFERENCES

[1] F. S. Board, "Artificial intelligence and machine learning in financial services," November, available at: http://www. fsb. org/2017/11/artificialintelligence-and-machine-learning-in-financialservice/(accessed 30th January, 2018), 2017.

[2] D. D. Hawley, J. D. Johnson, and D. Raina, "Artificial neural systems: A new tool for financial decision-making," Financial Analysts Journal, vol. 46, pp. 63-72, 1990.

[3] D. Cheng and P. Cirillo, "A reinforced urn process modeling of recovery rates and recovery times," Journal of Banking & Finance, vol. 96, pp. 1-17, 2018.

[4] K. Nian, T. F. Coleman, and Y. Li, "Learning minimum variance discrete hedging directly from the market," Quantitative Finance, vol. 18, pp. 1115-1128, 2018.

[5] T. Renault, "Intraday online investor sentiment and return patterns in the US stock market," Journal of Banking & Finance, vol. 84, pp. 25-40, 2017.

[6] K. Seeja and M. Zareapoor, "FraudMiner: A novel credit card fraud detection model based on frequent itemset mining," The Scientific World Journal, vol. 2014, 2014.

[7] O. B. Sezer and A. M. Ozbayoglu, "Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach," Applied Soft Computing, vol. 70, pp. 525-538, 2018.

[8] B. Weng, L. Lu, X. Wang, F. M. Megahed, and W. Martinez, "Predicting short-term stock prices using ensemble methods and online data sources," Expert Systems with Applications, vol. 112, pp. 258-273, 2018.

[9] C. X. Ling and C. Li, "Data mining for direct marketing: Problems and solutions," in Kdd, 1998, pp. 73-79.

[10] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques: Elsevier, 2011.

[11] E. I. Altman, G. Marco, and F. Varetto, "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)," Journal of banking & finance, vol. 18, pp. 505-529, 1994.

[12] F. Varetto, "Genetic algorithms applications in the analysis of insolvency risk," Journal of Banking & Finance, vol. 22, pp. 1421-1439, 1998.

[13] A. N. Kercheval and Y. Zhang, "Modelling high-frequency limit order book dynamics with support vector machines," Quantitative Finance, vol. 15, pp. 1315-1329, 2015.

[14] W. H. Beaver, "Financial ratios as predictors of failure," Journal of accounting research, pp. 71-111, 1966.

[15] P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques," Decision Support Systems, vol. 50, pp. 491-500, 2011.

[16] J. T. Wells, Occupational fraud and abuse: Obsidian Publishing Company, 1997.

[17] C. Spathis, M. Doumpos, and C. Zopounidis, "Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques," European Accounting Review, vol. 11, pp. 509-535, 2002.

[18] C.-C. Yeh, D.-J. Chi, and M.-F. Hsu, "A hybrid approach of DEA, rough set and support vector machines for business failure prediction," Expert Systems with Applications, vol. 37, pp. 1535-1541, 2010.

[19] J. W. Seifert, "Data mining and the search for security: Challenges for connecting the dots and databases," Government Information Quarterly, vol. 21, pp. 461-480, 2004.

[20] O. Aregbeyen, "The determinants of bank selection choices by customers: recent and extensive evidence from Nigeria," International Journal of Business and Social Science, vol. 2, pp. 276-288, 2011.

[21] K. O. Siddiqi, "Interrelations between service quality attributes, customer satisfaction and customer loyalty in the retail banking sector in Bangladesh," International Journal of Business and Management, vol. 6, p. 12, 2011.

[22] X. Hu, "A data mining approach for retailing bank customer attrition analysis," Applied Intelligence, vol. 22, pp. 47-60, 2005.

[23] K. Coussement and D. Van den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," Expert systems with applications, vol. 34, pp. 313-327, 2008.

[24] R. A. Soeini and K. V. Rodpysh, "Evaluations of Data Mining Methods in Order to Provide the Optimum Method for Customer Churn Prediction: Case Study Insurance Industry," in International Conference on Information and Computer Applications, 2012, pp. 290-297.

[25] T. O. Goonetilleke and H. Caldera, "Mining life insurance data for customer attrition analysis," Journal of Industrial and Intelligent Information, vol. 1, 2013.

[26] B. He, Y. Shi, Q. Wan, and X. Zhao, "Prediction of customer attrition of commercial banks based on SVM model," Procedia Computer Science, vol. 31, pp. 423-430, 2014.

[27] J.-H. Ahn, S.-P. Han, and Y.-S. Lee, "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry," Telecommunications policy, vol. 30, pp. 552-568, 2006.

[28] A. Sharma, D. Panigrahi, and P. Kumar, "A neural network based approach for predicting customer churn in cellular network services," arXiv preprint arXiv:1309.3945, 2013.

[29] E. W. Ngai, L. Xiu, and D. C. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," Expert systems with applications, vol. 36, pp. 2592-2602, 2009.

[30] P. S. Raju, D. V. R. Bai, and G. K. Chaitanya, "Data mining: techniques for enhancing customer relationship management in banking and retail industries," International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, pp. 2650-2657, 2014.

[31] H. Zhou, H.-f. Chai, and M.-l. Qiu, "Fraud detection within bankcard enrollment on mobile device based payment using machine learning," Frontiers of Information Technology & Electronic Engineering, vol. 19, pp. 1537-1545, 2018.

[32] K. Fu, D. Cheng, Y. Tu, and L. Zhang, "Credit card fraud detection using convolutional neural networks," in International Conference on Neural Information Processing, 2016, pp. 483-490.

[33] S. Chen, "Detection of fraudulent financial statements using the hybrid data mining approach," SpringerPlus, vol. 5, p. 89, 2016.

[34] A. K. Pujari, Data mining techniques: Universities press, 2001.

[35] J. R. Quinlan, "Simplifying decision trees," International journal of man-machine studies, vol. 27, pp. 221-234, 1987.

[36] S. Moro, R. Laureano, and P. Cortez, "Using data mining for bank direct marketing: An application of the crisp-dm methodology," in Proceedings of European Simulation and Modelling Conference-ESM'2011, 2011, pp. 117-121.

[37] E. Turban, R. Sharda, and D. Delen, "Decision Support and Business Intelligence Systems (required)," Google Scholar, 2010.

[38] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann, 2016.

[39] K. Chitra and B. Subashini, "Data mining techniques and its applications in banking sector," International Journal of Emerging Technology and Advanced Engineering, vol. 3, pp. 219-226, 2013.

[40] S. Ghosh, A. Hazra, B. Choudhury, P. Biswas, and A. Nag, "A Comparative Study to the Bank Market Prediction," in International Conference on Machine Learning and Data Mining in Pattern Recognition, 2018, pp. 259-268.

[41] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," European Journal of Operational Research, vol. 247, pp. 124-136, 2015.

[42] B. Wang, Y. Kong, Y. Zhang, D. Liu, and L. Ning, "Integration of Unsupervised and Supervised Machine Learning Algorithms for Credit Risk Assessment," Expert Systems with Applications, 2019.

[43] P. Cerchiello, P. Giudici, and G. Nicola, "Twitter data models for bank risk contagion," Neurocomputing, vol. 264, pp. 50-56, 2017.

[44] M. Islam and M. Habib, "A data mining approach to predict prospective business sectors for lending in retail banking using decision tree," arXiv preprint arXiv:1504.02018, 2015.

[45] Scikit-learn, "Scikit-learn," 2019.

[46] Kaggle, "Santander Customer Transaction Prediction - Can you identify who will make a transaction?," 2019.

[47] H. Khosravi and E. Kabir, "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties," Pattern recognition letters, vol. 28, pp. 1133-1141, 2007.

[48] C. Chatfield, "Exploratory data analysis," European journal of operational research, vol. 23, pp. 5-13, 1986.

[49] KNIME, "Dimensionality Reduction Techniques," 05/12/2015 2015.

[50] S. Raschka, "Linear Discriminant Analysis," Aug 3, 2014 2014.

[51] A. M. Martínez and A. C. Kak, "Pca versus lda," IEEE transactions on pattern analysis and machine intelligence, vol. 23, pp. 228-233, 2001.

[52] Medium, "Dimensionality Reduction(PCA and LDA)," 10 March, 2019 2019.

[53] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," Progress in Artificial Intelligence, vol. 5, pp. 221-232, 2016.

[54] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in Data mining and knowledge discovery handbook, ed: Springer, 2009, pp. 875-886.

[55] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," Journal of artificial intelligence research, vol. 61, pp. 863-905, 2018.

SBAG: A Hybrid Deep Learning Model for Large Scale Traffic Speed Prediction

Adnan Riaz[1]
School of Computer Science and Technology
Dalian University of Technology
Dalian, China

Mehak Khan[3]
Department of Computer Science and Technology
Harbin Institute of Technology
Harbin, China

Muhammad Nabeel[2]
School of Software Engineering
South China University of Technology
Guangzhou, China

Huma Jamil[4]
Department of Computer Science and Technology
PMAS Arid Agriculture University
Rawalpindi, Pakistan

***Abstract*—Intelligent Transportation System (ITS) is the fundamental requirement to an intelligent transport system. The proposed hybrid model Stacked Bidirectional LSTM and Attention-based GRU (SBAG) is used for predicting the large scale traffic speed. To capture bidirectional temporal dependencies and spatial features, BDLSTM and attention-based GRU are exploited. It is the first time in traffic speed prediction that bidirectional LSTM and attention-based GRU are exploited as a building block of network architecture to measure the backward dependencies of a network. We have also examined the behaviour of the attention layer in our proposed model. We compared the proposed model with state-of-the-art models e.g. Fully Convolutional Network, Gated Recurrent Unit, Long -short term Memory, Bidirectional Long-short term Memory and achieved superior performance in large scale traffic speed prediction.**

Keywords—Attention mechanism; large scale traffic prediction; Gated Recurrent Unit (GRU); Bidirectional Long-short term Memory (BiLSTM); Intelligent Transportation System (ITS)

I. Introduction

The performance of Intelligent Transportation System (ITS) applications principally depends on the quality of traffic data. Lately, with the increment of both traffic volume and traffic data, traffic speed perdition has become very important in the ITS. In the past decades, short term traffic prediction is under the eyes of researchers. Many researchers have proposed different approaches and netwroks in the the past decades which shows it has a long history and this issue is yet to resolve in case of accuracy about traffic speed predcition. To overcome and enhance the efficiency and accuracy of the traffic prediction, several approaches were proposed [1]. Many traffic amenities and applications are dependent on prediction accuracy.

Existing models coarsely divided into two categories, i.e. computational intelligence (CI) approaches and classical statistical methods indicated by previous literature [2][3][4]. The statistical methods were introduced at earlier stages when traffic data is limited and less complex but later with the advancement in traffic sensing technologies, arise of traffic data and computational power most of later work centers around computational intelligence approaches for traffic forecasting.

In general, regarding taking care of complex traffic forecasting problems [5], the computational approaches shattered the statistical methods like autoregressive integrated moving average (ARIMA) [6] in terms of capacity to catch nonlinear relationship and to deal with complex data. By the ascent of neural systems (NN) based methods, the full protentional of artificial intelligence was not subjugated in any case but many neural network-based models like Feed Forward Neural Network [7], Fuzzy Neural Network [8], Recurrent Neural Network [9], Gaussian Process [10] and hybrid Neural Network [11][12] are adopted for traffic forecasting problems. Recently, some hybrid architectures are proposed for traffic speed forecasting. Many factors influence on traffic forecasting, so single-component models are not suitable to complete the traffic prediction task. To make progress in the accuracy of traffic prediction, hybrid models are used in traffic speed prediction. In complex road network CapsNet [13] architecture proposed which replaced the max polling operation of CNN. To cope with the temporal evolution of traffic status, Recurrent Neural Networks (RNNs) models are specifically very appropriate because of the dynamic nature of transportation.

Structure of RNNs has internal memory with loops [14] that sequence data by maintaining a chain-like structure. However, RNNs are challenging to train during the backpropagating process because of the vanishing gradient problems, owing to the depth of the loop and chain-like structure. LSTMs addressed the aforementioned difficulties successfully. A spatial-temporal LSTM network, MapLSTM [15] for fine-grained traffic conditions. To predict traffic flow with big data, hybrid Deep Neural networks (DNN) [16] was proposed. Structural RNN (SRNN) proposed to deal with graph data of a road network [17].

LSTMs have the ability to deal with long term dependencies. In recent days, they have been gaining popularity in traffic forecasting because of a representative deep learning method handling sequence data. In the domain of transportation, the capability of LSTM is not fully utilized yet. To predict large-scaled transportation traffic, it is becoming a vital and challenging topic. In most of the existing studies, network-wide prediction achieved only, when for N nodes, the same number of N models were trained for a traffic network [18] because they use traffic data along with a corridor or sensor location. However, learning complex spatial-temporal features of network-wide traffic should be explored by only one model.

In terms of dependency in prediction problem, the LSTM process the information in the forward direction, so LSTM only process forward dependency[5]. There is highly possible that some useful information may not efficiently filter or passed, so to consider the backward dependencies is very important. The other reason to consider the backward dependency is periodicity in traffic data, because traffic conditions have strong regularity and periodicity[19]. As per the literature review, a few studies utilized backward dependency. To cover this gap, bidirectional LSTMs (BDLSTMs) architecture is adopted as a network structure component because it can handle both forward and backward dependencies. In a traffic network, the impact of downstream and upstream speed on any location cannot be ignored while predicting the large-scale traffic speed. Along a corridor, future speed values of a location are affected by past speed values of upstream and downstream locations that only use forward dependencies in time series data, shown from previous studies[19][12][20]. In spatial-temporal data, the learned feature will be more inclusive with both backward and forward dependencies.

In this paper, we proposed a hybrid deep learning model known as stacked bidirectional LSTM with attention GRU (SBAG) neural network for large scale traffic speed prediction. Our model achieved better performance with the comparison of state-of-the-art methods. We consider the traffic forecasting to a large scale traffic network rather than several adjacent locations or specific location along a corridor. We proposed a hybrid model considering the backward dependencies using Bidirectional LSTM to improve feature learning. We examined the behaviour of attention mechanism to make improvements in the proposed model.

The remainder of the paper is described as Section II Methodology, Section III Performance Evaluation, Section IV Conclusion.

II. Methodology

The component of the proposed model SBAG is detailed explained in this section.

A. Input Data

In this study, the proposed and the compared model takes the large-scale speed data as input, to take network-wide influences into account. When traffic jam propagates, it not only affects the nearby location but also far away locations in a whole network. In traffic speed prediction, the input data use a sequence of speed values along n historical time step at one location [2][18][21], denoted by a vector,

$$Y_T = [Y_{T-n}, Y_{T-(n-1)}, \dots, Y_{T-2}, Y_{T-1}] \tag{1}$$

Suppose the traffic network consists of P locations, and we need to predict the traffic speeds at time T using n historical time frames (steps), the input can be characterized as a speed data matrix,

$$Y_T^P = \begin{bmatrix} Y^1 \\ Y^2 \\ \vdots \\ Y^P \end{bmatrix} = \begin{bmatrix} Y_{T-n}^1 & Y_{T-n+1}^1 & \cdots & Y_{T-2}^1 & Y_{T-1}^1 \\ Y_{T-n}^2 & Y_{T-n+1}^2 & \ddots & Y_{T-2}^2 & Y_{T-1}^2 \\ \vdots & \vdots & & \vdots & \vdots \\ Y_{T-n}^P & Y_{T-n+1}^P & \cdots & Y_{T-2}^P & Y_{T-1}^P \end{bmatrix} \tag{2}$$

Where each element Y_T^P is the speed at 'pth' location and 'tth' time steps. To signify temporal attributes of speed data and streamline the expression of the equation, vector $Y_T^P = [Y_{T-n}, Y_{T-(n-1)}, \dots, Y_{T-2}, Y_{T-1}]$ represents the speed matrix where each element signifies 'P' locations speed values.

B. Bidirectional Long Short Term Memory (BDLSTMs)

The idea of using Bidirectional LSTMs comes from bidirectional RNN. The bidirectional LSTMs join two hidden-layers to the same output layer. Bidirectional LSTMs showed superiority in different fields over unidirectional e.g. speech-recognition [22], phoneme-classification [23]. The structure of Bidirectional LSTMs is shown in Fig. 1.

$$h_t = \sigma_h(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{3}$$

$$\hat{y}_t = \sigma_h(W_{hy}h_t + b_y) \tag{4}$$

$$\mathrm{f}_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{5}$$

$$\mathrm{i}_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{6}$$

$$\mathrm{o}_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{7}$$

$$\tilde{C}_t = \tan h\,(W_C x_t + U_C h_{t-1} + b_C) \tag{8}$$

Where, W_f W_i, W_o, W_C U_f, U_i, U_o, U_C. b_f, b_i, b_o, and b_C, are the weight matrices and bias vector parameter which need to be learned during training. σ_g is the gate activation function and hyperbolic tangent function being $\tan h$.

$$\hat{y}_t = \sigma(\overleftarrow{h}, \vec{h}) \tag{9}$$

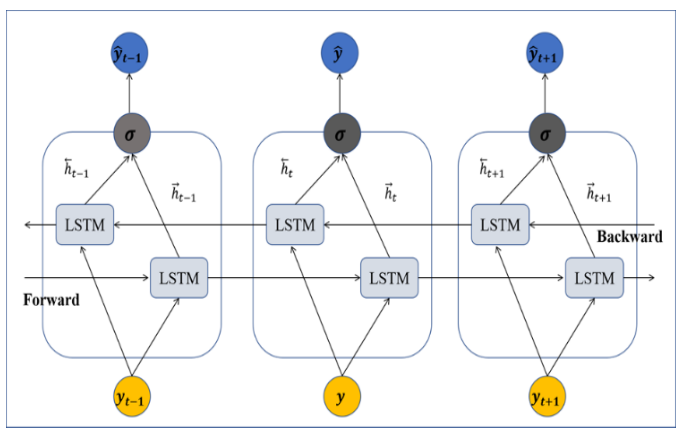


Fig. 1. Unfold Architecture of BDLSTM.

Where $\vec{h}$ and, $\overleftarrow{h}$ are the forward and backward layer output that iteratively calculated by using positive sequence inputs from time $T-n\ to\ T-1$ and vice versa, backward and forward layers outputs are calculated by eq. 3-8, $\hat{Y}_T$ is an output vector that can be generated by Bidirectional LSTM, where each element is calculated from the eq. 9. Where σ is an average function used to join the two output sequences.

C. GRU

GRU is a well-known variant based on LSTM proposed by Cho et al. [24]. GRU is simpler than LSTM because it has fewer parameters than LSTM and its performance is significant in some tasks. It consists of forget gate and input gate. It combines forget-gate and input-gate to an update-gate. In Fig. 2 GRU block diagram is shown. The memory cell of a GRU has four components that allow cells to access and save information for a longer time period. GRU calculate the hidden states by following equations:

$$z_t = \sigma\left(W^{(z)}.[h_{t-1}, x_t]\right) \quad (10)$$

$$r_t = \sigma\left(W^{(r)}.[h_{t-1}, x_t]\right) \quad (11)$$

$$\tilde{h}_t = tanh(W.[r_t * h_{t-1}, x_t\ ]) \quad (12)$$

$$h_t = (1-z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (13)$$

In the above equations,z_t and r_t are update and reset gate. σ is an activation function. $\tilde{h}_t$ is candidate activation function, h_t is the actual activation-function of the proposed GRU at time t.

D. Attention Mechanism

The hidden unit of GRU consists of an update and reset gate that captures dependencies of different timescales. In time series sequence, spatial-temporal dependency has not contributed equally, so attention mechanism is adopted in this paper with GRU to solve this problem [25]. The magnitude of the weights $\widehat{\propto_i}$ learned by the network signifies the importance of hidden states. We compute $\hat{r}$ as combination of all $\widehat{h_i}$, after using the attention mechanism.

$$\hat{r} = \sum_{i=1}^{N=1} \widehat{\propto}_i\ \hat{h}_i \quad (14)$$

At each time step, the hidden-state vector $\widehat{h_i}$ is the input of the attention layer. For this time step the attention weight $\widehat{w_i}$ can be calculated as

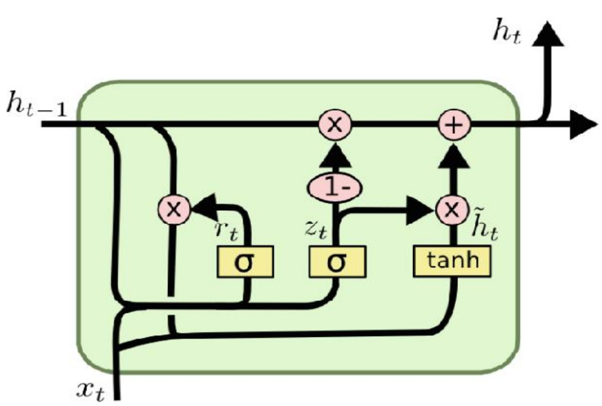


Fig. 2. A typical GRU Block Diagram.

$$\widehat{w}_i = \tan h\left(\widehat{h_i}\right) \quad (15)$$

$$\hat{b} = a^T m_i + b \quad (16)$$

$$\widehat{\propto_i} = \frac{\exp(\hat{b})}{\sum_k \exp(\hat{b})} \quad (17)$$

Where parameters of attention layer are a and b. At i_{th} time-step the attention layer output is formulated as:

$$\widehat{r_i} = \widehat{\propto}_i\ \hat{h}_i \quad (18)$$

E. BDLSTM-GRU Module (with Attention Mechanism) for Spatial-Temporal Correlation Features Learning

Existing studies demonstrated that LSTMs work effectively in sequence tasks. BDLSTMS has the power to process data in both ways backward and forward direction so we adopted BDLTSM as the first layer in our proposed model to capture spatial-temporal information while feeding input to the model during the feature learning process. The top layer of the model only required learned feature when predicting future speed values. We used GRU as the last layer of the architecture the output of the BDLSTM is fed into GRU as input We also utilized attention mechanism to enhance the capability of GRU to process large scale traffic speed prediction.

In this paper, we proposed a novel hybrid model stacked Bidirectional LSTM and attention GRU (SBAG) for the large scale traffic speed prediction. The proposed model takes spatial input and predicts traffic speed value for the next time step. Fig. 3 illustrates the architecture of the proposed model.

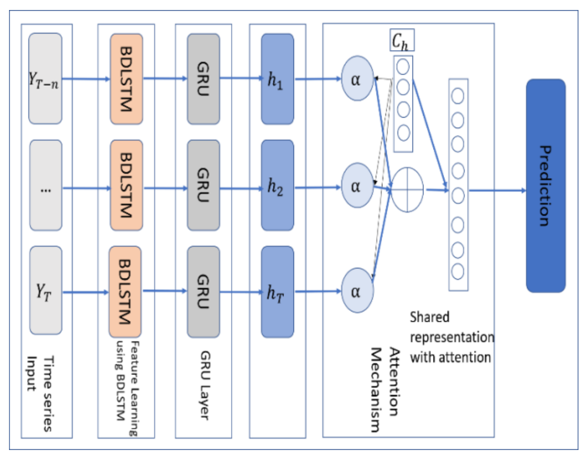


Fig. 3. Block Diagram of Proposed Model.

III. Performance Evaluation

A. Dataset Description

In this study, we used a publicly available dataset knows as a loop detector used by authors [26]. The dataset covers I-5, I-405, I-90, and SR-520 connected freeways; it has 5 minutes time step interval and 323 sensor stations and covers 5 minutes intervals over the entirety of 2015. Fig. 4 is the diagram of the loop detector dataset.

B. Experimental Setup

In input data Y_T^P, each sample is a 2-dimensional vector. The dimensions of input data are [n, P] = [10,323], based on model description. The time lag is set as 10.

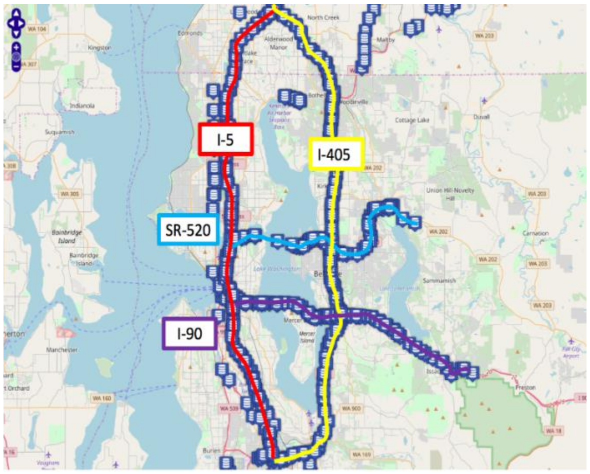


Fig. 4. Loop Detector Dataset.

C. Model Optimization

In training, mini-batch gradient descent is used. MSE is used as a loss function and RMSProp optimizer. To avoid overfitting, the early stopping mechanism is used.

$$loss = \frac{1}{N}\sum_{i=1}^{N}\left(h_{pred} - h_{true}\right)^2 \quad (19)$$

Where h_{pred} is the predicted results, h_{true} is the ground truth value and N denotes the number of training samples.

D. Evaluation Criteria

To evaluate the effectiveness of state-of-the-art models, Mean Squared Errors (MSE), Root Mean Squared Errors (RMSE) and R2 are calculated using the following equations:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(h_{pred} - h_{true}\right)^2 \quad (20)$$

$$RMSE = \sqrt{\sum_{N}^{1}\frac{\left(h_{pred} - h_{true}\right)^2}{N}} \quad (21)$$

$$R^2 = 1 - \frac{Explanied\ Variation}{Total\ Variation} \quad (22)$$

Where h_{pred} is the predicted results, h_{true} is the ground truth value and N denotes the number of training samples. Fig. 5 and Fig. 6 show the training and validation loss of different models compared in this study.

E. Comparison with State-of-the-Art Models

We compared the proposed model with state-of-the-art models to check the efficiency and effectiveness of the model. We compared the proposed model with Fully Convolutional Networks (FCN), Gated Recurrent Unit (GRU), Long-short Term Memory (LSTM), Long-short Term Memory with Deep Neural Network (LSTM-DNN) and Bidirectional Long-short Term Memory (BDLSTM). The performance comparison of different algorithms are demonstrated in Table I. The Means Squared Error and Root mean squared error of Fully convolutional network is 0.68 and 8.22 respectively. The performance of simple GRU is better than Fully Convolutional Network and the MSE and RMSE are reduced to 0.25 and 5.01 respectively. Because GRU cannot process long sequences and simple GRU is not a suitable choice for this problem. So, we compared the performance of LSTMs in this study. LSTM can address the short comings of GRU because of the gated structure of LSTM and the results are significantly better than GRU and FCN and error reduced to 0.1541 and 3.92. Furthermore, we added a DNN layer with LSTM and demonstrated that results are comparatively better, and error reduced to 0.1463 and 3.83. We also compared the performance of BDLSTM, and we concluded that the performance of BDLTSM is better than LSTM because BDLSTM process the sequence data both backward and forward and error reduces to 0.1408 and 3.75, respectively. The proposed model composed of BDLSTM and Attention-based GRU and its performance is superior over state-of-the-art models compared in this study and error reduces to 0.1371 and 3.70, respectively. Additionally, we also calculated the R^2 factor, which showed the performance of every model compared in this study and R^2 values of the proposed model are higher than other state of the art models compared which is 0.8620. In summary, we can come to know that the proposed model SBAG outperformed over FCN, GRU, LSTMS and BDLSTMs, as shown in Table I, respectively. Fig. 7 shows the error of models compared in this paper.

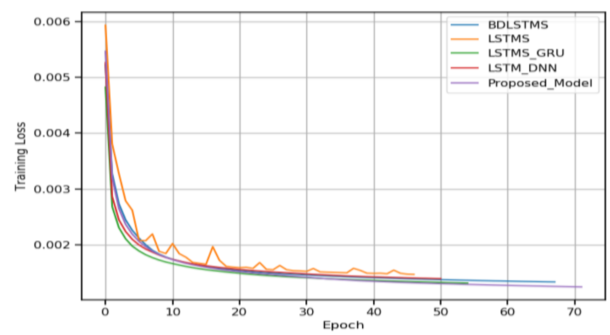


Fig. 5. Training Loss of different Models.

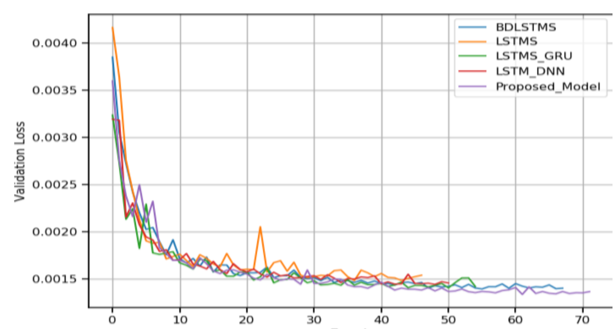


Fig. 6. Validation Loss of different Models.

TABLE. I. COMPARISON WITH STATE OF THE ART

Model	**MSE**	**RMSE**	**R^2**
FCN	0.6772	8.22	0.48
GRU	0.2512	5.01	0.7165
LSTM	0.1541	3.92	0.8550
LSTM-DNN	0.1463	3.83	0.8530
BiLSTM	0.1408	3.75	0.8548
(SBAG) Proposed Model	**0.1371**	**3.70**	**0.8620**

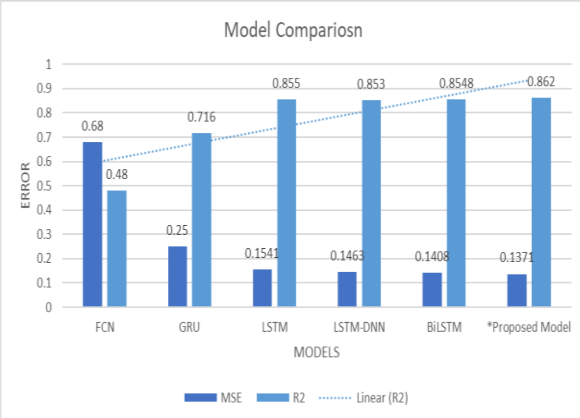


Fig. 7. Model Comparision Results.

IV. Conclusion

In this study, Stacked Bidirectional Long-Short Term Memory Attention-based Gated Recurrent Unit (SBAG) neural network is proposed for the large scale traffic speed prediction. The contributions and improvements focus on the three aspects: 1) Along a corridor, we consider the traffic forecasting to a large scale traffic network instead of specific location/several adjacent locations; 2) we proposed a hybrid model considering both backward and forward dependencies of large scale traffic data; 3) We considered the significance of attention layer that by adding attention layer with GRU, the performance of the proposed model significantly enhanced. Experiment results show that the SBAG is the best model for predicting large scale traffic speed. In comparison with GRU, LSTMs and BDLSTMs methods, Bidirectional LSTM and Attention-based GRU proven to be more competent to learn spatial-temporal features and the best model for large scale traffic speed prediction.

In the future, further improvements can be made based on the proposed study and will improve more significant towards the graph structure to interpret and learn the spatial features or to improve by hybridizing with GANs [27] to make further amendments.

References

[1] H. van Lint and C. van Hinsbergen, "Short-Term Traffic and Travel Time Prediction Models," Transp. Res. E-Circular, 2012.

[2] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," Transp. Res. Part C Emerg. Technol., vol. 54, pp. 187–197, May 2015.

[3] S. D. Khan, F. Porta, G. Vizzari, and S. Bandini, "Estimating speeds of pedestrians in real-world using computer vision," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2014.

[4] S. M. B. Muhammad Saqib, Sultan Daud Khan, "Vehicle Speed Estimation using Wireless Sensor Network," in INFOCOMP 2011 : The First International Conference on Advanced Communications and Computation, 2011.

[5] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," Transp. Res. Part C Emerg. Technol., 2011.

[6] Q. Ye, W. Y. Szeto, and S. C. Wong, "Short-term traffic speed forecasting based on data recorded at irregular intervals," IEEE Trans. Intell. Transp. Syst., 2012.

[7] D. Park and L. R. Rilett, "Forecasting freeway link travel times with a multilayer feedforward neural network," Comput. Civ. Infrastruct. Eng., 1999.

[8] H. Yin, S. C. Wong, J. Xu, and C. K. Wong, "Urban traffic flow prediction using a fuzzy-neural approach," Transp. Res. Part C Emerg. Technol., 2002.

[9] J. W. C. van Lint, S. P. Hoogendoorn, and H. J. van Zuylen, "Freeway Travel Time Prediction with State-Space Neural Networks: Modeling State-Space Dynamics with Recurrent Neural Networks," Transp. Res. Rec. J. Transp. Res. Board, 2007.

[10] P. Wang, Y. Kim, L. Vaci, H. Yang, and L. Mihaylova, "Short-Term Traffic Prediction with Vicinity Gaussian Process in the Presence of Missing Data," in 2018 Symposium on Sensor Data Fusion: Trends, Solutions, Applications, SDF 2018, 2018.

[11] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu, "Deep Learning: A Generic Approach for Extreme Condition Traffic Forecasting," in Proceedings of the 2017 SIAM International Conference on Data Mining, 2017.

[12] Z. Cui, R. Ke, and Y. Wang, "Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction," arXiv Prepr. arXiv1801.02143, 2018.

[13] Y. Kim, P. Wang, Y. Zhu, and L. Mihaylova, "A Capsule Network for Traffic Speed Prediction in Complex Road Networks," in 2018 Symposium on Sensor Data Fusion: Trends, Solutions, Applications, SDF 2018, 2018.

[14] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An Empirical Exploration of Recurrent Network Architectures," JMLR, 2015.

[15] X. Wei, J. Li, Q. Yuan, K. Chen, A. Zhou, and F. Yang, "Predicting Fine-Grained Traffic Conditions via Spatio-Temporal LSTM," Wirel. Commun. Mob. Comput., 2019.

[16] Y. Wu, H. Tan, L. Qin, B. Ran, and Z. Jiang, "A hybrid deep learning based traffic flow prediction method and its understanding," Transp. Res. Part C Emerg. Technol., 2018.

[17] Y. Kim, P. Wang, and L. Mihaylova, "Structural Recurrent Neural Network for Traffic Speed Prediction," arXiv Prepr. arXiv1902.06506, 2019.

[18] Y. Duan, Y. Lv, and F. Y. Wang, "Travel time prediction with LSTM neural network," in IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2016.

[19] X. Jiang and H. Adeli, "Wavelet packet-autocorrelation function method for traffic flow pattern analysis," Comput. Civ. Infrastruct. Eng., 2004.

[20] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting & Control. 2015.

[21] Y. Y. Chen, Y. Lv, Z. Li, and F. Y. Wang, "Long short-Term memory model for traffic congestion prediction with online open data," in IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2016.

[22] A. Graves, N. Jaitly, and A. R. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings, 2013.

[23] S. Xuan, H. Kanasugi, and R. Shibasaki, "DeepTransport: Prediction and simulation of human mobility and transportation mode at a citywide level," in IJCAI International Joint Conference on Artificial Intelligence, 2016.

[24] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," 2015.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," Microbes Infect., vol. 11, no. 3, pp. 367–373, Sep. 2014.

[26] Z. Cui, K. Henrickson, R. Ke, and Y. Wang, "Traffic Graph Convolutional Recurrent Neural Network: A Deep Learning Framework for Network-Scale Traffic Learning and Forecasting," 2018.

[27] M. Nabeel, A. Riaz, and W. Zhenyu, "Cas-GANs: An approach of dialogue policy learning based on GAN and RL techniques," Int. J. Adv. Comput. Sci. Appl., 2019.

Categorizing Attributes in Identifying Learning Style using Rough Set Theory

Dadang Syarif Sihabudin Sahid[1], Riswan Efendi[2], Emansa Hasri Putra[3], Muhammad Wahyudi[4]
Department of Information Technology, Politeknik Caltex Riau, Pekanbaru, Indonesia[1, 4]
Department of Mathematics, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia[2]
Department of Industrial Engineering, Politeknik Caltex Riau, Pekanbaru Indonesia[3]

***Abstract*—In a learning process, learning style becomes one crucial factor that should be considered. However, it is still challenging to determine the learning style of the student, especially in an online learning activity. Data-driven methods such as artificial intelligence and machine learning are the latest and popular approaches for predicting the learning style. However, these methods involve complex data and attributes. It makes it quite heavy in the computational process. On the other hand, the literate based driven approach has a limitation in inconsistency between results with the learning behavior. Combination, both approaches, gives a better accuracy level. However, it still leaves some issues such as ambiguity and a wide of range of attributes value. These issues can be reduced by finding the right approach and categorization of attributes. Rough set proposed the simple way that can compromise with the ambiguity, vague, and uncertainty. Rough set generated the rules that can be used for prediction or classification decision attributes. Yet, due to the method based on categorical data, it must be careful in determining the category of attributes. Hence, this research investigated several categorizing attributes in the identification learning style. The results showed that the approach gives a better prediction of the learning style. Different categories give different results in terms of accuracy level, number of eliminated data, number of eliminated attributes, and number of generated rules criteria. For decision making, it can be considered by balancing of these criteria.**

Keywords*—*Learning style; rough set; categorizing attributes; conditional attributes; decision attributes

I. Introduction

The shifting paradigm in the learning process from teacher-centered learning to students centered learning has changed the way of learning. Conventional ways of learning that emphasize one fits for all are no longer compatible with current conditions [1]. Each student has a unique and different way of learning. They have their own way and learning style. Moreover, the revolution of internet technology has provided various learning materials and media [2]. Technology attracts the student in a different way of learning. This situation has encouraged the development of learning models designed to follow students' personal needs in the form of learning personalization. In case of e-learning, the design of e-learning models that were initially technology-oriented and general in nature became more oriented to the needs, characteristics, situations, and conditions of students such as learning styles, prior knowledge, learning goals, cognitive abilities, learning interests, and motivation as parameters in learning [3]. Therefore, the identification of student learning styles is significant in the learning process.

The study related to the identification of learning styles is delivered in order to improve effectiveness and performance in learning [4]. However, the approach taken is still less efficient because it is done by conducting a series of questionnaires as well as an inconsistent result between the questionnaire and student behavior when the learning conducted. In general, learning style identification can be made through data-driven methods or literate based driven methods [5]. The data-driven method is conducted by transforming the questionnaire and using sample data sets to build a learner model. On the other hand, literate based-driven uses user behavior that provides instructional learning preferences when interacting with the learning environment [5]. Artificial intelligence and data mining methods are often used for the analysis of models based on a data-driven approach [6], [7]. On the other hand, a literate based driven approach uses a simple rule-based to compute the process of learning style models [8].

Both approaches have advantages and limitations. For example, using data-driven methods with sufficient data sets and appropriate methods are accurate enough to model learning styles. However, it is often encountered that is very complex with large enough data, so burdensome in the computing process. The efficiency advantage of a literate based driven approach is quite helpful in the computational process, although generally only suitable for modeling stationary and deterministic data. Whereas in learning style, there are often things that are dynamic, non-deterministic, and non-stationary [9].

Statistical modeling and individual machine learning are promising abilities for accurate predictive [10]. Integrating of two approaches using stochastic process and literate based driven has been conducted, but it still leaves some problems. This approach does not significantly distinguish learning styles. It has a similar distribution. So that it still raises ambiguity [11]. On the other hand, each attribute has a wide range of values. This requires an approach to be converted to a simpler range of values. Based on its characteristics, the rough set approach has the ability to resolve these issues. However, it needs to be further identified related to the process of categorizing appropriately for the existing attributes. Differences in determining categories can cause differences in the accuracy of predictions or classifications [12]. Therefore, this research will investigate several categorizing attributes and their effects on the level of accuracy in identifying learning styles. This paper is organized as follows: Introduction (Section I), Theoretical Background (Section II), Research Methodology (Section III), Results and Discussion (Section IV), and Conclusion (Section V).

II. Theoretical Background

A. Learning Style

Learning style is a characteristic of cognitive, affective, and psychomotor behavior, as an indicator that acts relatively stable for students to feel interconnected and react to the learning environment. Learning styles are learning habits that are preferred by learners [13]. In studies related to personalization e-learning, learning style plays an important role in a model of e-learning personalization [14][15]. This is to illustrate how the diversity of user conditions in learning results in different patterns of approach and learning preferences. Some are more interested in learning material in the form of text, video, audio, or pictures. In the meantime, maybe more interested in the way of presentation, such as in the form of concepts, examples, and other case studies. This shows that the learning process follows the needs of students in accordance with its learning style. So the learning style is one of the important components in the representation of learner e-learning personalization models [16].

B. Rough Set

Rough Set theory, firstly was introduced by Zdzislaw Pawlak in 1980, which was used to analyze of data classification in the form of information systems [17]. This theory uses a non-statistical data analysis approach. The purpose of the rough set analysis is to get a short estimate of the rule from a table or a data set. The results of the rough set analyses can be used in the process of data mining and knowledge discovery. Rough sets have been used widely in many fields such as medicine, pharmacology, economics, engineering, image processing, and decision analyses. The rough set is often used to modeling data with ambiguity, vague, and consist of uncertainty [18].

There are several important components contained in rough sets theory, namely: information systems and decision systems, indiscernibility relations, upper and lower approximations, discernibility matrix, data reduction, generated rules, and data prediction [19]. It can figure out information systems I=(U,Ω,Vq,fq) as follows [17][20]:

: Universe set;

Ω : Set of attributes;

$\Omega = C \cup D$, C is a finite set of condition attributes, and D is a finite set of decision attributes;

for each $q \in \Omega$, Vq is called the domain of q;

: an information function $fq : U \rightarrow Vq$;

C. Rough Set and its Application

Implementations of the rough set theory have been conducted in several studies. Bello and Verdegay presented the place of rough set in soft computing. The study combined the rough set with soft computing methods such as fuzzy logic, artificial neural network, and metaheuristics [21]. This hybridization approach succeeds in improving the performance of the system. Application rough set to identify behavior patterns of bank customer results 90% accuracy level. It is based on decision rule generation to predict the deposit nature of customers [18]. The rough set becomes effective tools for classifying 26 large scale construction projects in Iran and the other five countries. This classification is used to address the requirements and specifications of the construction project [22]. Korvin, et al. proposed the rough set theory to improve website performance by developing specific preloading strategy tuned to the needs of a web server. This approach was implemented due to the uncertainty of the internet user's behavior [23]. Another rough set application is used as one of the research methods to discover useful hidden patterns from fabric data to reduce the number of defective goods and increase overall quality. It is expected to improve the performance of manufacturing quality control activity and reduces productivity loss [24].

III. Research Methodology

This paper conducted several steps in order to achieve the research. The first step is data collection, then followed by categorizing conditional attributes, conducting rough set algorithm, identifying eliminated data, identifying eliminated conditional attributes, generated rules, and model evaluation.

A. Data Collection

In this research, learning style data was be obtained from an e-learning log server that involved 60 students who were taking IT Project Management Subjects. These students came from two classes. Although the data just have 60 records, but it has been taken repetitively every two weeks following the topic. They have been observed during ten weeks with five topics. Every topic provided four specific learning materials associated with the learning styles. Once the students visit a specific learning material, the counter will record duration visit (tMV, tMA, tMR, tMK). This first visit will also be recorded as a frequency of visit to specific learning material (fMV, fMA, fMR, fMK). If the student accesses the learning material at a different time, then the duration of the visit and the frequency of visit will also be added accumulatively. The data is recorded in an e-learning server log.

The conditional attributes consist of student learning behavior during interacting with the e-learning. The attributes are frequency visit and duration time of students when visiting specific learning material associated with a learning style. These conditional attributes can be shown in Table I. While the decision attributes consist of learning style based on VARK (Visual, Auditory, Read, Kinesthetic) that was introduced by Fleming [25]. This research used some learning material related to a specific learning style, as showed in Table II.

B. Categorizing Conditional Attributes

Data in conditional attributes are quantitative data. Frequency visit is measured by ordinal number, while duration time is measured in minutes. As required by rough set theory, the conditional attributes data should be converted to categorical data. In this research, the data is converted using categorizing criteria, as followed in Table III.

TABLE. I. CONDITIONAL ATTRIBUTES

No	Conditional Attributes	
	Name	***Remark***
1	tMV	Duration visit to visual learning material
2	tMA	Duration visit to auditory learning material
3	tMR	Duration visit to read learning material
4	tMK	Duration visit to kinesthetic learning material
5	fMV	Frequency visit to visual learning material
6	fMA	Frequency visit to auditory learning material
7	fMR	Frequency visit to read learning material
8	fMK	Frequency visit to kinesthetic learning material

TABLE. II. LEARNING MATERIAL

No	Learning Material	
	Type	***Associated Learning Style***
1	Video/Picture	Visual
2	Lecture (Monolog)	Auditory
3	Text (Slide, eBook)	Read
4	Instruction sheet, exercise	Kinesthetic

TABLE. III. CATEGORIZING CONDITIONAL ATTRIBUTES

Categorizing Conditional Attributes	
Number of Category	***Category***
2	High: Data ≥ Mean
	Low: Data < Mean
3	High: Data > Mean + SD
	Medium: Mean - SD ≤ Data ≤ Mean + SD
	Low: Data < Mean - SD
4	Very High: Data ≥ Q3
	High: Q2 ≤ Data < Q3
	Medium: Q1 ≤ Data < Q2
	Low: Data < Q1
5	Very High: Data ≥ Mean + 1.8SD
	High: Mean + 0.6SD ≤ Data < Mean + 1.8SD
	Medium: Mean - 0.6SD ≤ Data < Mean + 0.6SD
	Low: Mean – 1.8SD ≤ Data < Mean – 0.6SD
	Very Low: Data < Mean - 1.8SD

C. *Rough Set Algorithm*

This research conducted two main phases in implementing a rough set approach. These phases included eliminated unclassified data and conditional attributes reduction. The eliminated data follow the several steps [17].

- Provide information systems. In this case is provided by learning style data
- Conduct indiscernibility matrix
- Set approximation (upper and lower)
- Process elimination data
- Develop rough membership
- Dependency of attributes

Meanwhile, the reduction attributes follow this process:

Process 1 (Discernibility matrix)

- Create a discernibility shelter
- Read the entire Rough Sets data line after it has been reduced
- Read each column of the conditional attribute on the row read
- Create temporary shelter
- Compare the same conditional attribute column in the current row with the previous row
- If the value of the conditional attribute is different, then add the conditional name attribute to the temporary storage
- After the column reading is complete, compare the decision attribute in the current row with the previous row
- If the value decision attribute is not the same, then add temporary storage data to the discernibility storage

Process 2 (Boolean algebra)

- Create attribute reduction reservoir
- Read each group of conditional attribute data in the discernibility collection in sequence
- Compare the current conditional attribute group with the existing conditional attribute in the reduction attribute collection
- If there is the same conditional attribute, then the same conditional attribute becomes the value in the reduction attribute collection
- If not, then the conditional attribute group is now the value of the attribute reduction container
- The result of the conditional attribute reduction is the final value in the reduction reservoir

IV. RESULTS AND DISCUSSION

After conducting several steps of the research, there are some results that can be achieved in identifying learning styles using rough set theory. Categorizing conditional attribute gives different results in terms of the number of eliminated data, generated rules for prediction, and accuracy level. It becomes interesting for further discussion. For processing data, this research uses a self-developed application. The application has the capability for converting data from quantitative to categorical, eliminated data based on the rough set, generate the rule, and conduct evaluation through accuracy level. The following sections will be provided complete results from two categories. Other categories are served in the results summary due to the processes are similar.

A. Conversion Data to Categorical Data

The original learning style data consist of 60 data and eight attributes. The piece of data can be shown in Table IV.

Based on Table V, it calculated some variables to determine the threshold for categorizing conditional attributes. These variables included mean, standard deviation (SD), minimum value, maximum value, and quartile (Q1-Q3). The variable value of conditional attributes can be shown in Table V and Table VI.

Data in Table IV is converted using two categories according to Table III and variable value in Table V – Table VI. The conversion process of data can be shown in Fig. 1. The process is the formula to categorize conditional attributes become two categories: High and Low.

TABLE. IV. LEARNING STYLE DATA

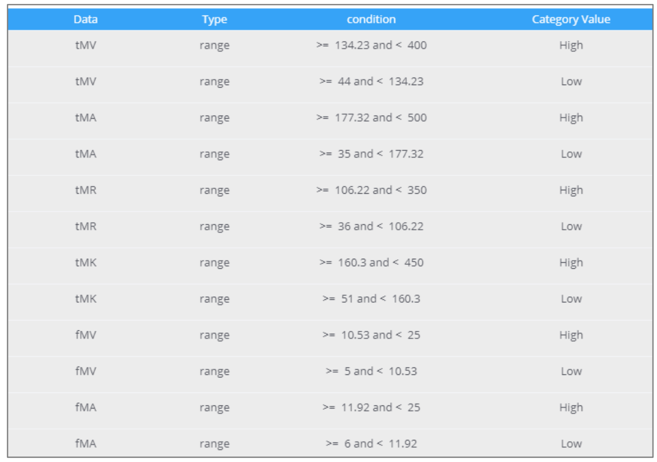
Student	***tMV***	***tMA***	***tMR***	**…**	***fMR***	***fMK***
S1	64	78	67	…	10	18
S2	76	170	80	…	10	16
S3	65	55	74	…	10	20
S4	286	48	82	…	8	5
S5	60	65	273	…	13	12
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
S56	86	80	279	…	15	12
S57	347	51	73	…	7	8
S58	44	448	51	…	6	9
S59	83	417	59	…	6	9
S60	315	138	42	…	9	8

TABLE. V. CONDITIONAL ATTRIBUTE VARIABLE (1)

Variable	**Conditional Attribute**			
	tMV	***tMA***	***tMR***	***tMK***
Mean	134.23	177.32	106.22	160.30
SD	107.68	140.54	81.89	122.54
Q1	66	70.75	65	62.75
Q2	77	85.5	74	91
Q3	223.75	283.75	88	233.75
Min	44	35	36	51
Max	387	448	343	413
Mean+SD	241.91	317.85	188.11	282.84
Mean-SD	26.55	36.78	24.33	37.76
Mean+1.8SD	328.06	430.28	253.62	380.88
Mean+0.6SD	198.84	261.64	155.35	233.83
Mean-06.SD	69.62	92.99	57.08	86.77
Mean-1.8SD	-59.59	-75.65	-41.18	-60.28

TABLE. VI. CONDITIONAL ATTRIBUTE VARIABLE (2)

Variable	**Conditional Attribute**			
	fMV	***fMA***	***fMR***	***fMK***
Mean	10.53	11.92	9.70	11.37
SD	5.56	4.92	4.01	4.96
Q1	7	8	7	7.75
Q2	8	9	9	9
Q3	14.25	17	10	15
Min	5	6	5	5
Max	23	23	24	23
Mean+SD	16.09	16.84	13.71	16.32
Mean-SD	4.97	7.00	5.69	6.41
Mean+1.8SD	20.54	20.77	16.92	20.29
Mean+0.6SD	13.87	14.87	12.11	14.34
Mean-0.6SD	7.20	8.96	7.29	8.39
Mean-1.8SD	0.53	3.06	2.48	2.44



Data	Type	condition	Category Value
tMV	range	>= 134.23 and < 400	High
tMV	range	>= 44 and < 134.23	Low
tMA	range	>= 177.32 and < 500	High
tMA	range	>= 35 and < 177.32	Low
tMR	range	>= 106.22 and < 350	High
tMR	range	>= 36 and < 106.22	Low
tMK	range	>= 160.3 and < 450	High
tMK	range	>= 51 and < 160.3	Low
fMV	range	>= 10.53 and < 25	High
fMV	range	>= 5 and < 10.53	Low
fMA	range	>= 11.92 and < 25	High
fMA	range	>= 6 and < 11.92	Low

Fig. 1. Data Conversion Process.

Categorizing conditional attributes into two categories involved mean, minimum, and maximum value of each conditional attributes. The conversion result of conditional attributes using two categories can be shown in Fig. 2. It is shown that the original data with variation value of mean and standard deviation was converted become two simple categorical values High and Low. It can be inferred that the generated rule as the basis of model development will have a simple rule. However, it can contain the issue due to two categories showing a wide range of value.

Student	tMV	tMA	tMR	tMK	fMV	fMA
S1	Low	Low	Low	High	Low	Low
S2	Low	Low	Low	High	Low	High
S3	Low	Low	Low	High	Low	Low
S4	High	Low	Low	Low	High	Low
S5	Low	Low	High	Low	Low	Low

1 2 3 4 5 6 ··· 5 Displaying 1 - 5 of 60 records

Fig. 2. Conversion Result.

B. Eliminated Data

As is mentioned in the previous section, data elimination is an important process in rough set theory [26]. By using the elimination process, the data becomes less than the original data, but it still gives the same results. Two steps in elimination are data and attribute reduction. These processes have a role in reducing redundant data or eliminated un-classified data. This is in line with the fact that many data sets are quite large, but not all of them can be used in forming models in decision making. It is possible that two or more conditional attributes have the same value but are inconsistent in their decision attributes. Consequently, the research on reduction attributes is an interesting and promising field. In this research, these reduction processes followed algorithms that have been figure out in the previous section.

Based on the algorithm, the screenshot code for eliminated data and attributes can be shown in Fig. 3 and Fig. 4. The data elimination algorithm is used to reduce the data which are identified in un-classified data. Explanation about un-classified data was delivered in previous section. Meanwhile, the conditional attributes elimination is used to reduce the attributes which are not affect to decision attributes. Both of algorithms were implemented using several categories as proposed in this research.

By using two categories, the eliminated result can be shown in Fig. 5. In this elimination process, 42 data have been successfully reduced. From the 60 original data, only 18 data were retained in building the prediction model. This information shows that with these two categories, many data are eliminated.

C. Generated Rules

In the previous section, it was shown that the elimination process carried out with two categories left 18 out of 60 original data. Based on this data, there are 13 generated rules that will be used as a basis for making predictions or classifications. If it is viewed from the computational process, these results show simplicity in the model. In the process of computing, this condition will make the process lighter. However, it is also important to look at the level of accuracy produced as a basis for further evaluation. The generated rule of the research by using two categories can be shown in Fig. 6.

The result of the generated rule also shows attributes that are used in predictions based on previous processes. These results also provide information on how complex the model is that it impacts the computational process. The number of simple rules with high accuracy can be used as an important reference in the selection for an optimal identification model.

D. Evaluation

The final stage of the learning style identification process is to evaluate the resulting level of accuracy. Previously, several criteria had been stated in the form of the amount of data eliminated and the number of rules generated in building the model. In this study, the resulting accuracy level was 96.67% as shown in Fig. 7. This level of accuracy is quite high, especially when compared to the data involved, and the rules are quite small.

```
// reduction process
dataRoughSetReduction()
__data = dataRealToKategori().get()
for (var i = 0; i < __data.length; i++) {
    __objData = __data[i]
    __newData = {}
    __strConAttr = ""
    __strConAttrDec = ""
    for (var header in __objData){
        __newData[header] = __objData[header]
        __dataHeader = dataHeader().filter({name: header}).get()
        if (__dataHeader[0]['type'] == 'conditional attribute'){
            __strConAttr += __objData[header].toLowerCase()+"-"
            __strConAttrDec += __objData[header].toLowerCase()+"-"
        }
        if (__dataHeader[0]['type'] == 'decission attribute'){
            __strConAttrDec += __objData[header].toLowerCase()+"-"
        }
    }
    __newData['__strConAttr'] = __strConAttr
    __newData['__strConAttrDec'] = __strConAttrDec
    dataRoughSetReduction.insert(__newData)
}
// reduction data double
__dataWithGroup = dataRoughSetReduction().distinct("__strConAttrDec")
for (var i = 0; i < __dataWithGroup.length; i++) {
    __dataFilter = dataRoughSetReduction().filter({__strConAttrDec: __dataWithGroup[i]}).get()
    if (__dataFilter.length > 1) {
        for (var j = 1; j < __dataFilter.length; j++) {
            thisId = __dataFilter[j]['___id']
            dataRoughSetReduction().filter({___id : thisId}).remove()
        }
    }
}
// reduction for boundary region
__dataWithGroup = dataRoughSetReduction().distinct("__strConAttr")
for (var i = 0; i < __dataWithGroup.length; i++) {
    __dataFilter = dataRoughSetReduction().filter({__strConAttr: __dataWithGroup[i]}).get()
    if (__dataFilter.length > 1) {
        for (var j = 0; j < __dataFilter.length; j++) {
            thisId = __dataFilter[j]['___id']
            dataRoughSetReduction().filter({___id : thisId}).remove()
        }
    }

}
```

Fig. 3. Screenshot Code for Data Elimination.

```
// stage reduction attribute
var DISCERNIBILITYMATRIX = []
for (var i = 0; i < dataRoughSets.length; i++) {
    row = dataRoughSets[i]
    for (var j = 0; j < i; j++) {
        compare = dataRoughSets[j]
        tempMAT = []
        for (var k = 0; k < compare.length; k++) {
            if (row[k] != compare[k]){
                tempMAT.push(header[k])
            } else {
                tempMAT.push('')
            }
        }
        if (decissionAttr[i] != decissionAttr[j]){
            DISCERNIBILITYMATRIX.push(tempMAT)
        }
    }
}
Result = []
for (var i = 0; i < DISCERNIBILITYMATRIX.length; i++) {
    _THIS = DISCERNIBILITYMATRIX[i]
    if (Result.length == 0){
        Result.push(_THIS)
    }
    else {
        _NEWRESULT = []
        _NOIRISAN  = 0
        for (var j = 0; j < Result.length; j++) {
            _RESULT   = Result[j]
            _NEWRESULT.push(_RESULT)
            _IRISAN   = __getIris(_RESULT)
            _FLAGPUSH = __cekIris(_IRISAN)
            __modifiedFlag(_FLAGPUSH)
        }
        if (_NOIRISAN == Result.length) {
            _NEWRESULT.push(_THIS)
        }
        Result = []
        Result = _NEWRESULT
    }
}
```

Fig. 4. Screenshot Code for Conditional Attribute Elimination.

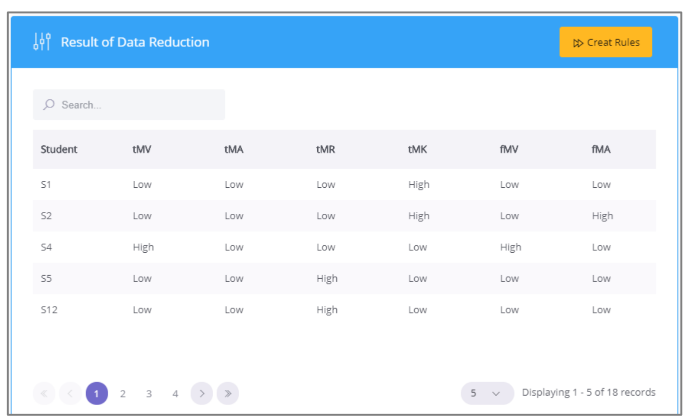


Fig. 5. Result of Data Elimination.

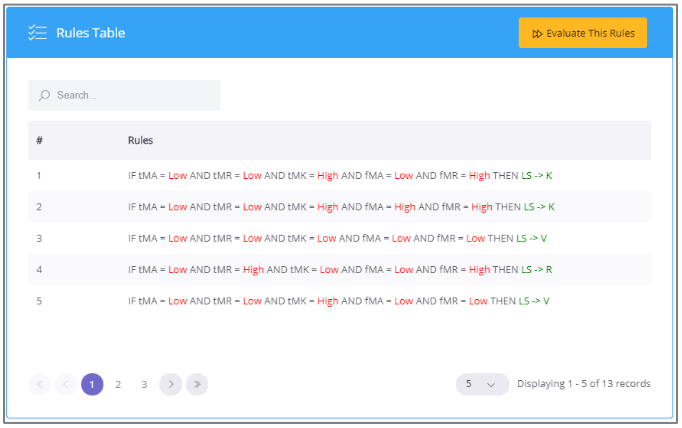


Fig. 6. Generated Rules.

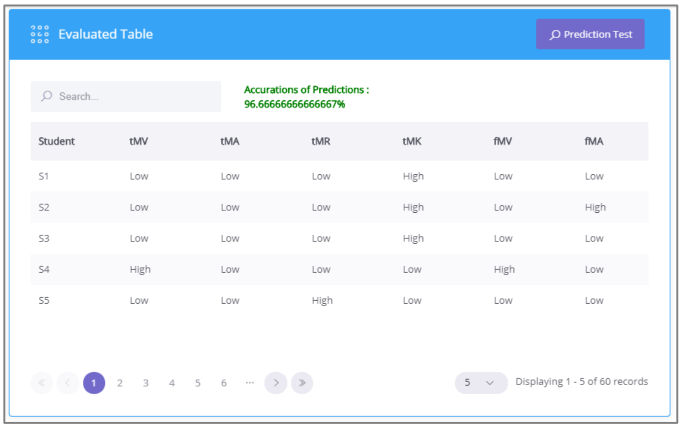


Fig. 7. Accuracy Level.

E. Result Summary

The results of categorizing the attributes with the two categories have been presented in the previous section. Furthermore, the results of the identification of learning styles with the number of categories 3, 4, and 5 are presented in Table VII.

Based on these results, category 3 produces 41 reduced data. This result is almost the same as the categorization of two categories. But with a higher number of generated rules, that is 16 rules. This three-category model, after evaluation, gives an accuracy rate of 93.33%. Category 4 provides 0 reduced data, 41 generated rules, and 100% accuracy. The results show that the number of rules produced provides maximum accuracy. Nevertheless, there are still issues related to processing computing with many rules. On the other hand, the process of reducing attributes with this category does not produce reduced data. Category 5 provides a maximum accuracy rate of 100%, 36 generated rules eight reduced data. These results are simpler in rule and reduced data compared to category 4. The results of learning style identification with the categorization of these attributes require further analysis as consideration for decision making.

TABLE. VII. RESULT SUMMARY FOR LEARNING STYLE

Number of Category	**Output**		
	Number of Eliminated Data	***Number of Generated Rule***	***Accuracy Level***
2	42	13	96.67%
3	41	16	93.33%
4	0	41	100%
5	8	36	100%

TABLE. VIII. RESULT SUMMARY FOR FLOW EXPERIENCE

Number of Category	**Output**		
	Number of Eliminated Data	***Number of Generated Rule***	***Accuracy Level***
2	39	4	74.74%
3	15	28	84.84%
4	7	61	92.92%
5	13	15	86.86%

As a comparison, the proposed method has been implemented in flow experience data set. This data set has 92 students. The similar result from the data set can be seen in Table VIII. The results show that the highest number of generated rules tend to give the highest accuracy level. Yet, a few generated rules such as in category 5, it still gives the high accuracy level. The balance between the criteria can be used as a foothold to choose the most optimal categorization to be used in the model.

V. CONCLUSION

Categorizing attributes in identifying learning styles has been presented in this research. Implementation of rough set theory with various categories in learning style gives high accuracy. The results provided alternatives that can be used in decision making process. It has a different number of rules and eliminated data. Some categories give the highest number of eliminated data with the lowest number of generated rules. But, the accuracy level is lower than the others. Other categories yield the highest accuracy level, but they have a minimum number of eliminated data and the number of generated rules. For decision-making purposes, it can be done by balancing three criteria: generated rule, number of eliminated data, and accuracy level. Each of the criteria has consequences, especially in the computational process. In this research, categorizing attributes used basic statistic descriptive parameter with normal distribution approach. It can be a limitation of the research. For future work, it can be investigated about the distribution of data before the categorizing process.

ACKNOWLEDGMENT

This research is funded by the Grant of Fundamental Research Scheme from the Ministry of Research, Technology, and Higher Education, Republic of Indonesia, 2019.

REFERENCES

[1] D. Syarif, S. Sahid, L. E. Nugroho, and P. I. Santosa, "A Model of Personalized Context-Aware E- learning Based on Psychological Experience," vol. X, no. X, 1843.

[2] A. Al Ahmad, "The Impact of Social Networks on Students ' Academic Achievement in Practical Programming Labs," vol. 10, no. 11, pp. 56–61, 2019.

[3] S. Graf, G. Yang, T. C. Liu, and Kinshuk, "Automatic , Global and Dynamic Student Modeling in a Ubiquitous Learning Environment," Knowl. Manag. E-Learning An Int. J., vol. 1, no. 1, pp. 18–35, 2009.

[4] T. U. Haider, A. K. Sinha, and B. D. Chaudhary, "An Investigation of relationship between learning styles and performance of learners," Int. J. Eng. Sci. Technol., vol. 2, no. 7, pp. 2813–2819, 2010.

[5] P. Q. Dung and A. M. Florea, "A literature-based method to automatically detect learning styles in learning management systems," Proc. 2nd Int. Conf. Web Intell. Min. Semant. - WIMS '12, p. 1, 2012.

[6] Ö. Şimşek, N. Atman, M. M. Inceoğlu, and Y. D. Arikan, "Diagnosis of learning styles based on active/reflective dimension of felder and Silverman's Learning Style Model in a learning management system," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6017 LNCS, pp. 544–555, 2010.

[7] N. Atman, M. M. Inceoğlu, and B. G. Aslan, "Learning styles diagnosis based on learner behaviors in web based learning," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5593 LNCS, no. PART 2, pp. 900–909, 2009.

[8] F. A. Dorça, L. V. Lima, M. A. Fernandes, and C. R. Lopes, "A Stochastic Approach for Automatic and Dynamic Modeling of Students ' Learning Styles in Adaptive Educational Systems," vol. 11, no. 2, pp. 191–212, 2012.

[9] F. A. Dorça, L. V. Lima, M. A. Fernandes, and C. R. Lopes, "Comparing strategies for modeling students learning styles through reinforcement learning in adaptive and intelligent educational systems: An experimental analysis," Expert Syst. Appl., vol. 40, no. 6, pp. 2092–2101, 2013.

[10] L. C. P. Velasco, D. L. L. Polestico, G. P. O. Macasieb, M. B. V Reyes, and F. B. V. Jr, "A Hybrid Model of Autoregressive Integrated Moving Average and Artificial Neural Network for Load Forecasting," vol. 10, no. 11, pp. 14–22, 2019.

[11] D. Syarif, S. Sahid, L. E. Nugroho, and P. I. Santosa, "Integrated Stochastic and Literate Based Driven Approaches in Learning Style Identification for Personalized E-Learning Purpose," vol. 7, no. 5, pp. 1708–1715, 2017.

[12] R. Efendi, "Rough-Regression for Categorical Data Prediction based on Case Study," 2019 Int. Conf. Artif. Intell. Inf. Technol., pp. 277–281, 2019.

[13] Willing K, Learning styles in adult migrant education. Adelaide: NCRC, 1988.

[14] H. M. Truong, "Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities," Comput. Human Behav., 2015.

[15] E. Özpolat and G. B. Akar, "Automatic detection of learning styles for an e-learning system," Comput. Educ., vol. 53, no. 2, pp. 355–367, 2009.

[16] M. Yarandi, A. R. H. Tawil, H. Jahankhani, and S. A. Hosseini, "Ontology-based learner modelling for supporting personalised e-Learning," Int. Conf. Interact. Mob. Comput. Aided Learn., pp. 113–118, 2012.

[17] P. Zdzislaw, "Rough Sets," Int. J. Comput. Inf. Sci., vol. 11, no. 5, pp. 341–356, 1982.

[18] L. Sumalatha, P. Uma Sankar, and B. Sujatha, "Rough set based decision rule generation to find behavioural patterns of customers," Sadhana - Acad. Proc. Eng. Sci., vol. 41, no. 9, pp. 985–991, 2016.

[19] S. Rissino and G. Lambert-Torres, "Rough Set Theory – Fundamental Concepts, Principals, Data Extraction, and Applications," 2009, pp. 35–58.

[20] R. Efendi, "Medipre : Medical Diagnosis Prediction Using Rough-Regression Approximation," pp. 1–5, 2018.

[21] R. Bello and J. L. Verdegay, "Rough sets in the Soft Computing environment," Inf. Sci. (Ny)., vol. 212, pp. 1–14, 2012.

[22] S. Akbari and M. Khanzadi, "Building a rough sets-based prediction model for classifying large-scale construction projects based on sustainable success index," 2018.

[23] G. Quirchmayr, "Information preloading strategies for e-government sites based on user ' s stated preference," vol. 20, no. 1, pp. 119–131, 2007.

[24] E. Kahya-o, "Enhancing product quality of a process," vol. 112, no. 8, pp. 1181–1200, 2012.

[25] N. Fleming, V.A.R.K Visual, Aural/Auditory, Read/Write, Kinesthetic. New Zealand: Bonwell Green Mountain Falls, 2006.

[26] J. Bai, K. Xia, Y. Lin, and P. Wu, "Attribute Reduction Based on Consistent Covering Rough Set and Its Application," vol. 2017, 2017.

Facial Emotion Recognition using Neighborhood Features

Abdulaziz Salamah Aljoloud[1], Habib Ullah[2], Adwan Alownie Alanazi[3]
College of Computer Science and Engineering
University of Ha'il, Ha'il, Saudi Arabia

***Abstract*—We present a new method for human facial emotions recognition. For this purpose, initially, we detect faces in the images by using the famous cascade classifiers. Subsequently, we then extract a localized regional descriptor (LRD) which represents the features of a face based on regional appearance encoding. The LRD formulates and models various spatial regional patterns based on the relationships between local areas themselves instead of considering only raw and unprocessed intensity features of an image. To classify facial emotions into various classes of facial emotions, we train a multiclass support vector machine (M-SVM) classifier which recognizes these emotions during the testing stage. Our proposed method takes into account robust features and is independent of gender and facial skin color for emotion recognition. Moreover, our method is illumination and orientation invariant. We assessed our method on two benchmark datasets and compared it with four reference methods. Our proposed method outperformed them considering both the datasets.**

Keywords—Haar features; feature integration; emotion recognition; face detection; localized features; multiclass SVM classifier

I. Introduction

Classification of emotion in different classes is a field of significant attention nowadays. The most important of this field is related to human facial emotion classification which is demonstrated as a chain procedure to recognize various human emotions via facial skin expressions (shown in Fig. 1), verbal expressions, different gesture and body movements, and different physiological signals measurement methods. The importance of people feelings in the research of latest technology gadgets is well-known. In today's world, the analysis and recognition of human emotion recognition has an extensive range of significance in wide majority of applications including machine learning based human-computer interaction, online automated tutoring systems, image and video retrieval, smart environments for health-care, and automated driver warning systems as narrated by Seyedehsamaneh et al. [1]. In addition to what has been mentioned above, facial emotion recognition plays very important role in finding various mental health conditions by doctors, psychiatrists and psychologists. In the past few decades, scientists and researchers from multidisciplinary fields have proposed different approaches and methods to identify emotions from facial features, speech signals, and many other sources. However, it is worth noticing that it is still a difficult issue in the field of machine learning, deep learning, computer vision, psychology, physiology due to the nature of its complexity. Facial recognition started nearly 80's [41]. Scientists and researchers agreed that facial expressions are the most influential part in recognizing human emotion. But, it is difficult to interpret human's emotion by utilizing facial expression characteristics due to the sensitivity to the external noises for example illumination conditions and dynamic head motion Kwang et al. [2]. Moreover, the final results for emotion classification based on facial expressions still need to be improved. For this purpose, different research investigations have been made and it is found out that the clue lies in the baseline or the backbone of most of the methods based on the initial step of face recognition. This fact was further investigated by Jiankang et al. [49]. They discovered that if a robust technique is used to detect faces, then the complexity of next steps can be reduced substantial and the effectiveness of these next steps improve significantly. Ray and Mishra [12] investigated EEG signals and on top of that they considered different techniques to measure the performance of emotion recognition capabilities.

To handle these problems, we introduce robust technique for human facial emotion classification into various states using facial features in the localized regions. Our proposed technique does not rely on the postulation of a specific gender or skin color of different human beings. The proposed technique is illumination and orientation invariant to prevail robustness to these changes. In fact, the proposed technique is characterized by the compact representation of spatial information as illustrated by Manisha et al. [3] that effectively combines human facial emotion features. We fuse the characterizations of both face detection and human facial emotion classification into a unique framework. The proposed technique follows the inspiration of investigating the local structure of facial image with a different technique of unification of localized features. It is important to mention here that the proposed technique is motivated by smaller computational overhead. This characteristic of the method makes this method very feasible to be placed in practice for any handheld device, for example, smart phones and other smart portable devices. The flow and complete process of our proposed technique is outlined in Fig. 2. We identified faces in the images using famous Haar features. Subsequently, we then formulate localized regional descriptor (LRD) and exploit multi-class SVM to classify different human facial emotion. Our contribution lies in the development of localized regional descriptor that motivates us further enhance the proposed method with experimental analysis from different aspects.

Fig. 1. Facial Emotions. different Human Emotion are Depicted from their Faces in the Provided Sample Images.

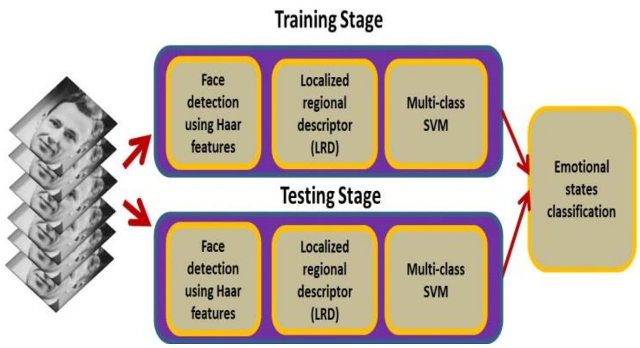


Fig. 2. Emotional States Classification using M-SVM Classifier. The Proposed Method Detect Faces in Video, and then Extract Localized Regional Descriptor (LRD) from the Detected Faces which are used to Train a Multiclass SVM During Training Stage. The Same Features are used to Classify Facial Emotions into different States During the Testing Stage.

In the rest of the paper, we present literature reviews in Section II and our proposed method in Section III for classification of facial emotions into seven different classes. Results are presented in Section IV, discussion is presented in Section V, and conclusion is presented in Section VI.

II. Literature Review

We provide details of literature review in this section. We have partitioned state-of-the-art techniques into 3 parts to explain human facial emotion classification considering speech signals, physiological signals measurements, and human facial expressions based recognition methods.

One dimensional signal, namely, speech which is a complicated signal providing a lot details, for instance, about the data to be communicated, speaker, language, region, and emotions. Therefore, we want to mention that speech processing is a significant field in digital signal processing and it presents a number of different applications including human computer interfaces using machine learning techniques, telecommunication between peer, assistive technologies for health-care, and security and safety associated with different places of people gatherings. The sound and speech/acoustic properties of the speech signal represent feature and the procedure through which some data is extracted from the speech signal and this is called feature extraction as introduced by Likitha et al. [4] and they utilized Mel Frequency Cepstral Coefficient (MFCC) method for human facial emotion recognition through speech signals representing different properties. Lotfidereshgi et al. [5] introduced an algorithm that uses the speech signal directly from the provided data through various speech collection devices. Therefore their technique fuses the robustness of the traditional source filter model of human speech generation with those of the currently presented liquid state machine (LSM) which is also called as biologically-inspired spiking neural network (SNN). Tzirakis et al. [6] presented a technique consisting of a Convolutional Neural Network (CNN). This model formulates features from the unprocessed signal, and concatenates them together to present them to a 2-layer Long Short- Term Memory (LSTM) network. Taking into account speech emotion from multiple sources i.e., in multiple speech emotion, the rate of identifying emotion will be decreased due to the expansion of emotional confusion. To fix this issue, Sun et al. [7] presented a speech emotion recognition technique considering the decision tree support vector machine (SVM) algorithm with Fisher feature selection bottom-up approach. Liu et al. [8] introduced a speech emotion recognition technique considering an enhanced version of brain emotional learning (BEL) algorithm, which is motivated by the emotional processing procedure of the limbic system in the brain of human beings. The outcome results of BEL algorithm is affected and improperly adapted by the reinforcement learning rule. Moreover, human emotions classification considering speech signals suffer from the unavailability of information and features because they don't provide improved interaction between human and machine in the form of a computer. To enhance the robustness of speech signals information itself, still a very large amount of technical space should be completed and addressed by the researchers in the same field.

Now we consider different category where emotions are classified using signal measurement procedure. For instance, physiological signals measurements are engendered by the physiological process of human beings, e.g., heart-beat rate (electrocardiogram or ECG/EKG signal of brain in the human), respiratory rate of human and content (capnogram), skin conductance (electro thermal activity or EDA signal on the body), muscle current (electromyography or EMG signal taken via different hardware sources available in the market), brain electrical activity (electroencephalography or EEG signal that can be measured using different electrodes on human skull). The aforementioned ways of signals collection help in finding emotion of human beings due to various mental and physical activities. For instance, Ferdinando et al. [9] used LDA technique (Linear Discriminant Analysis feature method), NCA (Neighbourhood Components Analysis feature method), and MCML (Maximally Collapsing Metric Learning for feature assessment) for the supervised monitoring and decreasing of different features in human emotion recognition based on ECG signals collected via electrodes. Kanjo et al. [10] presented a technique that removes the requirements for manual feature extraction by using multiple learning methods, for example, a hybrid method considering a deep model namely Convolutional Neural Network and another deep model, namely, Long Short-term Memory Recurrent Neural Network (CNN-LSTM) on the unprocessed sensor information based on phones and wearable devices easily available in the marked. Nakisa et al. [11] fixed the problem related with the high-dimensionality of EEG signals by presenting an algorithm to effectively search for the optimal subset of EEG features in

EEG signals. For this purpose, they used evolutionary computation (EC) methods. Moreover, taking into account signal pre-processing and emotion classification, their technique divides a huge set of emotions and combines extra features. Ray et al. (2019) introduced a method by using computational intelligence algorithm e.g., discrete wavelet transform and Bionic Wavelet Transform (BWT) for the evaluation of EEG signals Ullah et al. [13]. Jirayucharoensak et al. [14] investigated the usage of a deep learning network (DLN) to find out undiscovered feature correlation between input signals from various sources. The DLN is used with a stacked auto encoder using hierarchical feature learning technique. It is worth mentioning that the physiological signals measurement based techniques for human emotion classification face several issues as illustrated by Egon et al. [15]. These issues are obtrusiveness of physiological sensors, unreliability of physiological sensors, for example, due to movement artifacts of multiple reasons, not fixed bodily position, changing air temperature, and varying humidity. In addition to that, these signals have many-to-many relationship issues; that is, multiple physiological signals can partially serve as indicators for multiple conventional biometric features of human emotions. These signals also present varying time windows where measurements could differ.

Now we will provide details of methods based on facial emotion recognition aspects. Facial expression based emotions classification moves the next level the fluency of the environment, accuracy and genuineness of interaction taking place in the surroundings, especially to demonstrate human-computer interaction complications as illustrated by Rota et al. [16] in his method related to particle groupings. To take into account these considerations, both scientists and researchers from the community are contributing important efforts to facial expression based emotion classification techniques and the literature is increasing with the passage of time. Jain et al. [17] introduced an algorithm based on advance and latest Deep Convolutional Neural Networks (DNNs) that is made of various layers performing different functions and deep residual blocks to achieve different tasks of interest. Wang et al. [18] proposed a technique considering stationary wavelet entropy to discover robust features, and used a single hidden layer feed forward neural network as the classifier for facial expression classification. Jaya method is presented to block the training of the classifier fall into local optimum regions that would ultimately compromise the overall performance. Yan et al. [19] introduced a novel and robust discriminative multi-metric learning approach for facial expression classification in multiple video. Orientation feature descriptors from many directions for each face video are discovered to illustrate facial appearance and motion data from dynamic aspects. These metrics driven by multiple features are subsequently learned with these extracted multiple features in a unified fashion to use complementary and discriminative data for emotion classification. Sun et al. [20] introduced a multi-channel deep neural network that learns and puts together the spatial-temporal descriptors for facial expressions identification in static frames. The important concept of the algorithm is to discover and collect optical flow from the difference among the peak expression face frame and the neutral face frame as the temporal data of a specific facial expression, and consider the grey-level frame of peak expression face as the spatial data. A Deep Spatial-Temporal feature Fusion neural Network is investigated to collect the performance of the deep feature extraction and combination from the frames and images. Lopes et al. [21] introduced a robust algorithm for facial expression identification that uses a unification of Convolutional Neural Network and some novel pre-processing factors for the same purpose. Chen et al. [22] proposed a robust method to handle the key challenge of face motions by considering a robust set of features namely Histogram of Oriented Gradients from three perpendicular planes to collect features associated with textures from video data. For the consideration and utilization of facial appearance variations, a robust geometric feature Ullah et al. [23] is introduced from a novel transformation of facial landmarks. Discovering the strengths of facial features based emotions classification techniques, people in the field paid attention to facial expression based emotion classification techniques for handheld smart devices including mobiles. To this end, smart mobiles and smart wrist watches are fully equipped with different types of sensors, for instance, accelerometer, gyroscope, fingerprint Sensor, heart rate sensor, and microphone. Alshamsi et al. [24] investigated a method driven by sensor technology and cloud computing for identification of emotion in both speech and facial expression. Hossain et al. [25] introduced a framework that puts together the strengths of emotion-aware big data and cloud technology towards 5G. In fact, they fused together facial and verbal descriptors to introduce a bimodal technique for big data emotion classification. Grünerbl et al. [26] presented a method considering smartphone sensors for the identification of depressive and panic mental states and recognize state variations of people targeted by bipolar disorder disease. Sneha et al. [27] introduced the textual content of the message and user typing behaviour to make a model that easily divides the future instances. Hossain et al. [28] introduced a method in which Bandlet transform is used on the face areas, and the resultant subband is partitioned into non-overlapping sections. Additionally, a local binary pattern is investigated for each section. The Kruskal-Wallis feature selection is used to choose the most discriminative bins of the fused histograms, which are provided to Gaussian mixture model-based classifier to find different human emotion. Sokolov et al. [29] presented a cross-platform system for human emotion identification. Their system is based on convolutional neural network. Their system can effectively identify human emotions on arousal-valence level of measurement. Lee et al. [42] proposed deep networks for context-aware emotion recognition that consider both human facial expression and context data in a combined fashion. Mao et al. [43] introduced three HMM based frameworks and compared throughout the current paper. Han et al. [44] investigated and summarized the ideas and categories, techniques and applications of transfer learning briefly, and studies the combination of transfer learning and deep learning, and the application of speech emotion recognition. Borra et al. [45] presented an attendance system using partial facial recognition. Nhuong et al. [46] propose an algorithm for feature extraction for the purpose of face recognition. Imen et al. [47] introduce sequence kernels for emotion recognition. Erfana et al. [48] present a survey about the emotion intelligence of different algorithms in the field.

The literature is very limited due to the associated challenges of developing a reliable technique with low computational requirements. The aforementioned methods require huge computational powers since most of them are based on deep models. These methods are modelled for very narrow and specific emotions and they are not extendable easily to consider other emotional states. Therefore, we propose an efficient method for emotions classification into a set of different states using facial features. Our method is independent of gender class, skin colour, illumination changes, and face orientations. Our proposed method presents compact representation of spatial information Verma et al. [3] that effectively encodes emotion information. We integrate the strengths of both face detection and emotion classification into a unified model. Additionally, our method is driven by low computational complexity. Therefore, it can be implemented easily on any handheld device including smart phones.

III. Proposed Method

Feature modeling for facial emotion classification has been an active area in the fields of image processing and computer vision. The motivation for fast face modeling for realistic facial recognition and classification has led scientists to discover different model-based methods. The techniques in the literature for facial expression modeling and recognition differ in various aspects depending on the application under observation, computing efficiency, type of sensors, cost, and required accuracy. Some researchers proposed 3D generic face deformation for smartphone applications, where they use a single image to adapt the generic model to the face in the video frame captured via smartphones. Different methods can be adapted facial features extraction from video frames. Some researchers use stereo to model face features using differential geometry. However, this kind of technique requires prior knowledge about the shape of the surfaces of the face and its differential geometry for accurate performance. Parallel stereo images can also be used that rely on manually selected corresponding feature points to compute the rotation and translation matrices that are used to fit the model to the computed feature points. For facial emotion recognition we model features which are illumination and orientation invariants. For classification we use multiclass SVM which is a powerful and accurate classifier. Multiclass SVM presents good performance on many problems including non-linear problems. Due to the classification strengths of multiclass SVM, our method avoids both overfitting and underfitting. Multiclass SVM renders good performance by training it even with small samples. Considering our proposed features, this makes the classifier ideal for different personality traits, and high segmented facial expression. The multiclass SVM presents generalization capability; therefore, our proposed method can handle unseen data. The generalization capability of our method is determined by complexity and training of the multiclass SVM.

Face detection considering Haar feature-based cascade classifiers is a famous face detection model Aguilar et al. [30] and Viola et al. [31] due to its simplicity and robustness. Inspired by the mode, where we train a cascade function considering ground truth faces with their labels. In fact, the model entails a lot of positive labels for faces and negative labels for non-faces to train the classifier. Subsequently, we extract Haar features which resemble convolutional kernel. Each feature is a single value calculated by subtracting sum of pixels under a rectangle from sum of pixels under a different rectangle considering a video frame under observation. Due to different rectangles, we exploit different sizes and locations of each kernel to obtain a lot of features. For this purpose, the concept of integral image is exploited.

$$\Phi(x,y) = \sum_{x'\leq x, y'\leq y} \Gamma(x',y')\xi(x,y) = \xi(x,y-1) + \xi(x-1,y) + \Gamma(x,y) \quad (1)$$

$$\Phi(x,y) = \Phi(y-1) + \Phi(x-1) + \xi(x,y)$$

Where Φ is the integral image and Γ(x', y') is the original image. ξ(x, y) is the cumulative row sum. The integral image can be obtained in one pass over the original image. Additionally, we explore Adaboost model to filter out irrelevant features. To remove irrelevant features, we consider each and every feature on all the training images. For each feature, we investigate the optimal threshold which will classify faces and non-faces. We choose the features with smallest error rate since these features classify the faces and non-faces in optimized way. In the beginning, each image is rendered an equal weight. After each classification, we increase the weights of misclassified images and repeat the same procedure. We then calculate new error rates and new weights. We found that in each video frame and image, significant section consists of irrelevant areas. Therefore, if part of a window does not contain face, we remove it. To consider the concept of Cascade of Classifiers is modeled, which fuse the features into different stages of classifiers and use them one-by-one instead of applying all the features on a window. We remove the widow if it does not qualify the first stage. Therefore, we do not explore the remaining features. If the window qualifies the first stage, we apply the second stage of features and continue the procedure. A window qualifying all stages is a face region.

We than extract localized regional descriptor (LRD) which represents the features of a face based on localized appearance encoding. The LRD formulates different pattern based on the relationships between local areas themselves instead of considering only intensity information. For appearance information, we use localized regions in numerous directions and scales to compute regional patterns. We find the correspondence between localized areas by using the extrema on appearance magnitudes. We want to efficiently summarize the local structures of face by using each pixel as center pixel in a region under observation. Considering a detected face, for a center pixel Δc and neighboring pixels Δn (n=1,2,..,8), we compute the pattern number (ω) as,

$$(\omega)m,n = \sum_{n=0}^{M-1} 2^n xy\Xi 1(\Delta n - \Delta c) \quad (2)$$

$$\Xi 1 = \begin{cases} 1, & if\ \ (\Delta n - \Delta c) < 0 \\ 0, & Otherwise \end{cases} \quad (3)$$

where M and N are the radius of neighbors and number of neighbors for the pattern number. After calculating the ω of face, histogram is computed as formulated in the equation,

$$\Upsilon 1(1) = \sum_{x=1}^{M} \sum_{y=1}^{N} \Xi 2(\omega x, y, 1): 1 \in [0, 2^M - 1] \quad (4)$$

Research deanship at the University of Ha'il, Saudi Arabia.

$$\Xi 2(a,b) = \begin{cases} 1, & if \quad a = b \\ 0, & Otherwise \end{cases} \quad (5)$$

Relationship between regions in terms of these pixels has been used, and a pattern number is assigned. We model histogram to represent the face in the form of LRD. For regional pixels Δ n and a center pixel Δ c, LRD can be formulated as,

$$\eta_1^n = \begin{cases} \eta_1^n = \Delta_8 - \Delta_n . \eta_2^n = \Delta_{n+1} - \Delta_n, \; for \;\; n = 1 \\ \Delta_{n-1} - \Delta_n, \eta_2^n = \Delta_{n+1} - \Delta_{n,}, \forall, n = 1,2, \ldots .8 \\ \eta_1^n = \Delta_{n-1} - \Delta_n, \eta_2^n = \Delta_1 - \Delta_n, for, n = 8 \end{cases} \quad (6)$$

We find the difference of each region with two other regions in ηn1 and ηn2. Considering these two differences, we assign a pattern number to each region,

$$\Xi_3(\eta_1^n - \eta_2^n) = \begin{cases} 1, if \;\; \eta_1^n \geq 0, and, \eta_2^n \geq 0 \\ 1, \;\; if \;\; \eta_1^n < 0, and, \eta_2^n < 0 \\ 0, \;\; if \;\; \eta_1^n \geq, and, \;\; \eta_2^n < 0 \\ 0, \;\; if \;\; \eta_1^n < 0, \;\; and, \eta_2^n \geq 0 \end{cases} \quad (7)$$

For the central pixel Δc, LRD can be found using the above numbers and the histogram for LRD map can be calculated in the equations,

$$LRD(\Delta c) = \sum_{n=1}^{8} 2^{n-1} x \Xi_3(\eta_1^n - \eta_2^n) \quad (8)$$

$$\Upsilon_2(LRD) = \sum_{x=1}^{M} \sum_{y=1}^{N} \Xi_2 \, (LRD_{x,y,}1) : 1 \in [0, 2^8 -]$$

The LRD represents robust features which are calculated by extracting the relationship among local regions by considering them mutually. The LRD finds the relationship of local regions with central region. In the proposed method, face detection and LRD are fused as they complete each other on the basis of characteristics they represent individually.

To classify LRD features into various classes, we use the M-SVM classifier Liu et al. [32] and Du et al. [33]. The M-SVM consists of different parameters which are a combination of different predictors. The M-SVM classifier takes the input features, classifies them with every set of parameters in the classifier, and provides the class label that obtained the majority of votes. The classifier is trained with the same parameters considering the training sets which are produced from the original training set using the bootstrap process. For each training set, the classifier identifies the same number of features as in the original set. The features are chosen with replacement. It means that some features will be taken more than once and some will be ignored. At each iteration of the algorithm, the classifier does not use all the variables to compute the best split, but a unpredictable subset of them. With each set of parameters a new subset is generated. The M-SVM classifier does not require any performance estimation process, such as cross-validation or bootstrap, or a separate test set to get an approximation of the training error. In fact, the error is calculated internally during the training. In fact, in machine learning, M-SVMs are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training samples, each labeled as associated to one or the other of two categories, an SVM classifier sets up a model that assigns new samples to one class or the other, making it a non-probabilistic binary linear classifier. An M-SVM classifier is a representation of the samples as points in space, mapped so that the samples of the separate classes are isolated by a clear gap that is as wide as possible. New samples are then mapped into that same space and predicted to associate to a class based on the side of the gap on which they fall. M-SVM can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

IV. Results

For experimental evaluation, we consider static facial expressions in the wild (SFEW) 2.0 Dhall et al [34] and Dhall et al. [35] dataset and real world affective faces (RAF) dataset Li et al. [36]. The dataset namely static facial expressions in the wild (SFEW) has been collected by choosing frames from AFEW part of the collection which is popular among the community of facial emotion recognition. The database presets a lot of images and frames representing unconstrained facial expressions, varied head poses, various age ranges, different occlusions, various focus, different resolution of face and close to real world illumination in both the background and foreground. These chosen facial frames are takes from AFEW sequences and labeled based on the label of the sequence. In summary, SFEW consists of many images and that have been marked for six different facial expressions including angry, disgust, fear, happy, sad, surprise and the neutral class and was labeled by two independent participants. Similarly, real-world affective faces database is very big facial expression dataset consisting of very diverse facial frames downloaded from the Internet. Based on different annotation technique, each individual frame has been independently marked by a huge number of participants. Frames in this dataset are of great varieties that include changes in age, gender and ethnicity, head poses, lighting conditions, occlusions, and post-processing operations. This dataset has large aforementioned diversities considering different factors, large quantities, and rich annotations. Additionally, we perform comparison with many state-of-the-art methods and reported the results in term of both confusion matrices and total accuracies. We consider seven facial emotion classes including sad, happy, angry, disgust, fear, neutral, and surprise.

We compare our proposed method with four reference methods over two datasets. These reference methods include implicit fusion model Han et al. [37], biorthogonal model Dong et al. [38], higher order model Ali et al. [39], and bioinspired model Vivek et al. [40]. The comparison results are listed in Table I in term of total accuracies. Our proposed method achieved promising results and performed better than four reference methods. Our method still has some limitations. For example, we did not exploit geometric features. Our method is applicable to treat and diagnosis patients with emotion issues. It is worth noticing that the anger facial expression is tense emotional outcome when the human considers that his/her personal limits are violated. Persons in this kind of emotion generally take the gestures including intense stare with eyes wide open, output uncomfortable sounds, bare the teeth, and attempt to physically seem larger. The staring with eyes wide open is a significant hint for computers to recognize anger considering other facial emotions. There are also other face related elements including V-shape eyebrows, wrinkled nose, narrowed eyes, and forwarded jaws. All these important elements help to recognize anger emotion.

In the facial expression, happiness indicates an emotional state of joy. In this emotional state, the reader can find that the forehead muscle relaxes and the eyebrows are pulled up slowly. Apart from that, both the wrinkled outer corners of eyes and pulled up lip corners represent unique representation. In fact, the neutral facial emotion relaxes the muscles of the face and other facial emotions all need to use extensive muscles of face. The other six facial emotions in the datasets are more extreme.

We have also provided the confusion matrix for SFEW dataset in Table II. As can be seen, our proposed method presents encouraging results regarding the facial emotions.

We have also provided the confusion matrix for RAF dataset in Table III. As can be seen, our proposed method presents encouraging results regarding the facial emotions.

A great diversity of approaches has been proposed to solve the problem of facial emotion recognition. However, most of them are designed to work for specific emotion, where different representations of structures and appearance are analyzed with different models. In this paper we consider spatial properties of faces considering different emotions. The facial emotions are complex spatial representation with unexpected appearance or spatial patterns. For facial emotion recognition, we propose a novel method where we compute localized regional descriptor from the face images. Considering these facial emotions, we design a set of robust features combined into a unified LRD descriptor. For compact encoding of spatial patterns in these faces, we explore regional pixels which represent distinguish spatial patterns of faces. In fact, localized regional features are mid-level characteristics to fuse the distance between low-level and high-level features for capturing facial emotions. For classification, we exploited M-SVM which is a set of supervised learning methods used for classification and regression. Provided a set of training samples, the SVM classifier builds a model that finds the class of new unseen samples. This classifier is very significant in both machine-learning and data-mining curriculums and is frequently used by researchers. Besides, its utilization spans to a wide variety of applied research fields including but not limited to neuroscience, text categorization, and finance. The effectiveness of M-SVMs classification tasks in a wide variety of fields, such as text or image processing and medical informatics, has inspired researchers to do research on the execution performance and scalability of the training phase of serial versions of the algorithm. Since we describe a facial emotion from the view of a set of features, our method can be widely exploited in different applications. What's more, our modeling does not limit the type of features or the type of scenes, which helps us to extend the proposed technique to broader research fields. Experimental results demonstrated that our proposed approach is effective for the detection of various facial emotions.

TABLE. I. TOTAL ACCURACIES ARE PRESENTED FOR THE REFERENCE METHODS AND OUR PROPOSED METHOD CONSIDERING BOTH THE DATASETS NAMELY SFEW AND RAF

Methods	**SFEW Total Accuracy**	**RAF Total Accuracy**
Han et al. [37]	56.4	55.7
Zhang et al. [38]	54.2	56.6
Ali et al. [39]	49.8	52.7
Vivek et al. [40]	53.5	54.9
Prop. method	58.3	59.2

TABLE. II. CONFUSION MATRIX FOR SFEW DATASET IS PROVIDED WHERE ENCOURAGING PERFORMANCE OF OUR PROPOSED METHOD IS SHOWN

	Angry	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
Angry	.54	.05	.06	.05	.04	.10	.02
Disgust	.04	.65	.20	.16	.01	.01	.03
Fear	.31	.10	.52	.01	.01	.04	.03
Happy	.04	.11	.01	.50	.01	.3	.02
Neutral	.09	.03	.10	.02	.63	.19	.05
Sad	.02	.12	.01	.07	.05	.51	.11
Surprise	.01	.02	.10	.01	.06	.13	.66

TABLE. III. CONFUSION MATRIX FOR RAF DATASET IS PROVIDED WHERE ENCOURAGING PERFORMANCE OF OUR PROPOSED METHOD IS SHOWN

	Angry	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
Angry	.56	.04	.15	.03	.17	.06	.01
Disgust	.13	.58	.01	.01	.02	.15	.11
Fear	.00	.01	.62	.03	.03	.20	.12
Happy	.07	.04	.05	.50	.01	.04	.30
Neutral	.03	.01	.02	.13	.66	.06	.10
Sad	.07	.00	.20	.02	.01	.70	.01
Surprise	.10	.08	.15	.01	.03	.05	.57

V. DISCUSSION

We have presented a new method for facial emotion recognition based image processing and computer vision techniques. It is worth mentioning here that many methods have proposed previously for the same problem as we discussed in the literature review. However, those methods suffer from various problems ranging from limited datasets to limited metrics for the purpose of evolutions. Moreover, our proposed method is invariant to different key challenges as we mentioned in the introduction section. We carried out detail experimental analysis on two benchmark datasets which are considered very challenging for the same problem in the community. Thanks for localized feature descriptor that proved that our method is enriched with robustness to deal with the difficult problem of facial emotion recognition. In the experimental assessment, we used two performance metrics i.e., total accuracy and confusion matrix. Our method showed very promising results considering both aforementioned datasets and performance metrics. In fact, our work can be further extended with many machine learning and deep learning approaches. However, these advance learning approaches required huge amount of data to process during the training stage. Therefore, we keep it our next step in the future.

VI. Conclusion

We explore a new method for facial emotion classification into seven different states. For this purpose, we detect faces and extract localized regional descriptor (LRD) based on the relationships between neighboring regions. To classify facial emotions into seven different classes, we train a multi-class SVM classifier which recognizes these emotions during the testing stage. We evaluated our method on two benchmark datasets and compared it with four reference methods that show that we outperformed them.

In our future work, we would like to consider publicly available datasets as well as we will collect our own datasets in order to have huge amount of data. Then we will explore a deep learning model for the same problem. A deep learning method will address the weaknesses associated with our method including the usage of limited datasets and the consideration of limited number human emotions for the purpose of classification.

Acknowledgment

This work was supported by the deanship of scientific research, University of Ha'il, Saudi Arabia [BA-1912].

References

[1] S. Shojaeilangari, W.Y. Yau, K. Nandakumar, J. Li, and E. Khwang Teoh (2015), "Robust representation and recognition of facial emotions using extreme sparse learning," IEEE Transactions on Image Processing, vol. 24, no. 7, pp. 2140–2152, 2015.

[2] K. E. Ko and K. B. Sim (2010), "Development of a facial emotion recognition method based on combining aam with dbn," in Cyberworlds (CW), 2010 International Conference on. IEEE, 2010, pp. 87–91.

[3] M. Verma and B. Raman (2018), "Local neighborhood difference pattern: A new feature descriptor for natural and texture image retrieval," Multimedia Tools and Applications, vol. 77, no. 10, pp. 11843–11866, 2018.

[4] M.S. Likitha, S. R. R Gupta, K. Hasitha, and A. U. Raju (2017), "Speech based human emotion recognition using mfcc," in Wireless Communications, Signal Processing and Networking (WiSPNET), 2017 International Conference on. IEEE, 2017, pp. 2257–2260.

[5] R. Lotfidereshgi and P. Gournay (2017), "Biologically inspired speech emotion recognition," in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 5135–5139.

[6] P. Tzirakis, J. Zhang, and B. W. Schuller (2018), "End-to-end speech emotion recognition using deep neural networks," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5089–5093.

[7] S., S. Fu, and F. Wang (2019), "Decision tree svm model with fisher feature selection for speech emotion recognition," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2019, no. 1, pp. 2, 2019.

[8] Z.T. Liu, Q. Xie, M. Wu, W. H. Cao, Y. Mei, and J.W. Mao (2018), "Speech emotion recognition based on an improved brain emotion learning model," Neurocomputing, 2018.

[9] H. Ferdinando, T. Seppänen, and E. Alasaarela (2017), "Enhancing emotion recognition from ecg signals using supervised dimensionality reduction.," in ICPRAM, 2017, pp. 112–118.

[10] E. Kanjo, E. M. G Younis, and C. Siang Ang (2019), "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," Information Fusion, vol. 49, pp. 46–56, 2019.

[11] B. Nakisa, M. N. Rastgoo, D. Tjondronegoro, and V. Chandran (2018), "Evolutionary computation algorithms for feature selection of eeg-based emotion recognition using mobile sensors," Expert Systems with Applications, vol. 93, pp. 143–155, 2018.

[12] P. Ray and D. P. Mishra (2019), "Analysis of eeg signals for emotion recognition using different computational intelligence techniques," in Applications of Artificial Intelligence Techniques in Engineering, pp. 527– 536. Springer, 2019.

[13] H. Ullah, M. Uzair, A. Mahmood, M. Ullah, S. D. Khan, and F. A. Cheikh (2019), "Internal emotion classification using eeg signal with sparse discriminative ensemble," IEEE Access, vol. 7, pp. 40144–40153, 2019.

[14] S. Jirayucharoensak, S. P. Ngum, and P. Israsena (2014), "Eegbased emotion recognition using deep learning network with principal component based covariate shift adaptation," The ScientificWorld Journal, vol. 2014, 2014.

[15] E. L. V. D. Broek and M. Spitters (2013), "Physiological signals: The next generation authentication and identification methods!?," in Intelligence and Security Informatics Conference (EISIC), 2013 European. IEEE, 2013, pp. 159–162.

[16] P. Rota, H. Ullah, N. Conci, N. Sebe, and F. G. B. De Natale (2013), "Particles cross-influence for entity grouping," in 21st European Signal Processing Conference (EUSIPCO 2013). IEEE, 2013, pp. 1–5.

[17] D. K. Jain, P. Shamsolmoali, and P. Sehdev (2019), "Extended deep neural network for facial emotion recognition," Pattern Recognition Letters, 2019.

[18] S. H. Wang, P. Phillips, Z. C. Dong, and Y. D. Zhang (2018), "Intelligent facial emotion recognition based on stationary wavelet entropy and jaya algorithm," Neurocomputing, vol. 272, pp. 668–676, 2018.

[19] H. Yan (2018), "Collaborative discriminative multi-metric learning for facial expression recognition in video," Pattern Recognition, vol. 75, pp. 33–40, 2018.

[20] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han (2017), "Deep spatialtemporal feature fusion for facial expression recognition in static images," Pattern Recognition Letters, 2017.

[21] A. T. Lopes, E. de Aguiar, A. F De Souza, and T. O. Santos (2017), "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," Pattern Recognition, vol. 61, pp. 610–628, 2017.

[22] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," IEEE Transactions on Affective Computing, vol. 9, no. 1, pp. 38–50, 2018.

[23] H. Ullah, A. B. Altamimi, M. Uzair, and M. Ullah (2018), "Anomalous entities detection and localization in pedestrian flows," Neurocomputing, vol. 290, pp. 74–86, 2018.

[24] H. Alshamsi, V. Kepuska, H. Alshamsi, and H. Meng (2018), "Automated facial expression and speech emotion recognition app development on smart phones using cloud computing," in 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE, 2018, pp. 730–738.

[25] M. S. Hossain, G. Muhammad, M. F. Alhamid, B. Song, and K. Al-Mutib (2016), "Audio-visual emotion recognition using big data towards 5g," Mobile Networks and Applications, vol. 21, no. 5, pp. 753–763, 2016.

[26] A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, S. Oehler, G. Tröster, O. Mayora, C. Haring, and P. Lukowicz (2015), "Smartphone-based recognition of states and state changes in bipolar disorder patients," IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 1, pp. 140–148, 2015.

[27] H.R Sneha, M. Rafi, M.V. M. Kumar, L. Thomas, and B. Annappa (2017), "Smartphone based emotion recognition and classification," in Electrical, Computer and Communication Technologies (ICECCT), 2017 Second International Conference on. IEEE, 2017, pp. 1–7.

[28] M. S. Hossain and G. Muhammad (2017), "An emotion recognition system for mobile applications," IEEE Access, vol. 5, pp. 2281–2287, 2017.

[29] D. Sokolov and M. Patkin (2018), “Real-time emotion recognition on mobile devices,” in Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on. IEEE, 2018, pp. 787–787.

[30] W. G. Aguilar, M. A. Luna, J. F. Moya, V. Abad, H. Parra, and H. Ruiz (2017), “Pedestrian detection for uavs using cascade classifiers with meanshift,” in Semantic Computing (ICSC), 2017 IEEE 11th International Conference on. IEEE, 2017, pp. 509–514.

[31] P. Viola and M. Jones (2001), “Rapid object detection using a boosted cascade of simple features,” in Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. IEEE, 2001, vol. 1, pp. I–I.

[32] L. Shuangyin, L. Xu, Q. Li, X. Zhao, and D. Li. "Fault Diagnosis of Water Quality Monitoring Devices Based on Multiclass Support Vector Machines and Rule-Based Decision Trees." IEEE Access 6 (2018): 22184-22195.

[33] D. Shichang, C. Liu, and L. Xi. "A selective multiclass support vector machine ensemble classifier for engineering surface classification using high definition metrology." Journal of Manufacturing Science and Engineering 137, no. 1 (2015): 011003.

[34] D. Abhinav, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. "Emotion recognition in the wild challenge 2014: Baseline, data and protocol." In Proceedings of the 16th international conference on multimodal interaction, pp. 461-466. ACM, 2014.

[35] D. Abhinav, R. Goecke, S. Lucey, and T. Gedeon. "Collecting large, richly annotated facial-expression databases from movies." IEEE multimedia 19, no. 3 (2012): 34-41.

[36] S. Li, W. Deng, and J. P. Du (2017), “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2852–2861.

[37] J. Han, Z. Zhang, Z. Ren, and B. Schuller (2019), “Implicit fusion by joint audiovisual training for emotion recognition in mono modality,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5861–5865.

[38] Y. D. Zhang, Z. J. Yang, H. M. Lu, X. X. Zhou, P. Phillips, Q. M. Liu, and S. H. Wang (2016), “Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation,” IEEE Access, vol. 4, pp. 8375–8385, 2016.

[39] H. Ali, M. Hariharan, S. Yaacob, and A. H. Adom (2015), “Facial emotion recognition based on higher-order spectra using support vector machines,” Journal of Medical Imaging and Health Informatics, vol. 5, no. 6, pp. 1272–1277, 2015.

[40] T. V. Vivek and G. R. M. Reddy (2015), “A hybrid bioinspired algorithm for facial emotion recognition using cso-ga-pso-svm,” in 2015 Fifth International Conference on Communication Systems and Network Technologies. IEEE, 2015, pp. 472–477.

[41] B. Arthur L. "The neuropsychology of facial recognition." American Psychologist 35, no. 2 (1980): 176.

[42] L. Jiyoung, S. Kim, S. Kim, J. Park, and K. Sohn. "Context-aware emotion recognition networks." In Proceedings of the IEEE International Conference on Computer Vision, pp. 10143-10152. 2019.

[43] M. Shuiyang, D. Tao, G. Zhang, P. C. Ching, and T. Lee. "Revisiting Hidden Markov Models for Speech Emotion Recognition." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6715-6719. IEEE, 2019.

[44] H. Zhijie, H. Zhao, and R. Wang. "Transfer Learning for Speech Emotion Recognition." In 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), pp. 96-99. IEEE, 2019.

[45] S. Borra, K. J. Nazare, S. V. Raju, and N. Dey. "Attendance recording system using partial face recognition algorithm." In Intelligent techniques in signal processing for multimedia security, pp. 293-319. Springer, Cham, 2017.

[46] L. D. Nhuong, G. N. Nguyen, L. V. Chung, and N. Dey. "MMAS Algorithm for Features Selection Using 1D-DWT for Video-Based Face Recognition in the Online Video Contextual Advertisement User-Oriented System." Journal of Global Information Management (JGIM) 25, no. 4 (2017): 103-124.

[47] T. Imen, M. S. Bouhlel, and N. Dey. "Discrete and continuous emotion recognition using sequence kernels." International Journal of Intelligent Engineering Informatics 5, no. 3 (2017): 194-205.

[48] Z. S. Erfana, A. M. Khan, A. K. Srivastava, N. G. Nguyen, and N. Dey. "A study of the state of the art in synthetic emotional intelligence in affective computing." International Journal of Synthetic Emotions (IJSE) 7, no. 1 (2016): 1-12.

[49] D. Jiankang, J. Guo, N. Xue, and S. Zafeiriou. "Arcface: Additive angular margin loss for deep face recognition." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4690-4699. 2019.

A New Solution to Protect Encryption Keys when Encrypting Database at the Application Level

Karim El bouchti[1], Soumia Ziti[2], Fouzia Omary[3], Nassim Kharmoum[4]
Department of Computer Science, IPPS Team
Faculty of Sciences, Mohammed V University in Rabat
Morocco

***Abstract*—Encrypting databases at the application level (client level) is one of the most effective ways to secure data. This strategy of data security has the advantage of resisting attacks performed by the database administrators. Although the data and encryption keys will be necessarily stored in the clear on the client level, which implies a problem of trust viz-a-viz the client since it is not always a trusted site. The client can attack encryption keys at any time. In this work, we will propose an original solution that protects encryption keys against internal attacks when implementing database encryption at the application level. The principle of our solution is to transform the encryption keys defined in the application files into other keys considered as the real keys, for encryption and decryption of the database, by using the protection functions stored within the database server. Our proposed solution is considered as an effective way to secure keys, especially if the server is a trusted site. The solution implementation results displayed better protection of encryption keys and an efficient process of data encryption /decryption. In fact, any malicious attempt performed by the client to hold encryption keys from the application level cannot be succeeded since the real values of keys are not defined on it.**

Keywords—Database encryption; encryption key protection model; database encryption keys protection; data security

I. Introduction

Today, Database (DB) security is considered as one of the significant challenges in the computer world. Recently, it has been the subject of several debates and studies by experts and researchers in the data security field. The main purpose is to protect sensitive data against unavailability, leakage, and modification face to attackers' threats [9], [21], [22].

With the rapid development of technology, the attack scenarios on the DBs have become easy to realize. Currently, a penetration to a DB using SQL injection techniques becomes fast and quite simple thanks to specialized tools [10]. Different threats may come from various sources, some from trusted DB users, others from external ones, and some attacks are performed by the DB administrators [1], [2], [3]. Data theft is a dangerous attack. However, attacks compromising data integrity can generate heavier consequences; they are hardly detectable as the theft of data. For this reason, the implementation of three security levels is necessary: 1 / - Physical security, 2 / - Operating system security, 3 / - Database Management System (DBMS) security [15], [18].

The implementation of these security levels is useful, especially the access control mechanism implemented at the DBMS level. It is considered as the first defense line against unauthorized access [14], [23]. However, this mechanism is not sufficient, it protects only against attacks coming from outside the information system, and some limits can be generated as it has been explained in [11], [16]. Besides the access control mechanism, a security approach based on DB encryption can play an important role; it can be performed on three levels: DB level, hard disk level, and application level [7].

Encrypting DB at the application level consists of delegating encryption and decryption to applications that are connected to it. The DB server processes and manipulates only encrypted data, and the encryption keys are not implemented on the server. This solution provides strong protection of the encryption keys from threats performed by the administrator. However, the keys must necessarily be stored either on the applications or in a place where the applications are managed (application server for example). This approach is limited as the application user (or the application server administrator) are not always trusted sites, they can attack encryption keys at any time, and decrypt all sensitive data without leaving any traces [16].

It is mentioned in [8] that DBs confront several internal threats, especially those coming from legitimate users of the system. These threats can be realized in collaboration with a malicious administrator, another legitimate user, or with a malicious attacker outside the information system [1], [12]. Excessive privilege abuse, legitimate privilege abuse, and elevation of privilege are some models of internal attacks that compromise encryption keys. A legitimate DB user can exploit the privileges one’s to attack encryption keys directly. The user can also exploit uncorrected DB vulnerabilities or DB configuration errors to access illegally to the DB encryption keys location. A mistrust application administrator may reveal the encryption keys to an external attacker in order to attack the DB outside the perimeter of the information system [8], [13]. Trusted users are considered a serious threat to the security of encryption keys when encryption is performed at the application level. Otherwise, the internal attacks are considered dangerous not only on DBs but on Big Data platforms like Hadoop. For instance, the authors of [24] argued that an effective attack launched from the compromised node could degrade the data processing performance of the cluster. Then, they exposed an effective schema that might mitigate the risk of this attack and keep the cluster running efficiently. Also, the authors of [25] have developed a solution called ROVER, which is an efficient and verifiable Erasure Coding based Storage (ECS) for Big Data platforms. They showed that this solution implementation has good robustness and effectiveness against attacks from compromised nodes.

In order to protect encryption keys within applications, several solutions have been proposed by researchers. Among the most important works, we highlight the work proposed by Ding et al. [19]. They have proposed a new data encryption model implemented at the application level that ensures the confidentiality of sensitive data. It is based on a new method using keys chain to protect encryption keys. The authors of [20] have proposed a special concept called "Encryption as a service" to encrypt DB. Its main goal is to outsource the encryption system outside applications as an encryption service provider unit independently of applications users and the DB server. The protection of encryption keys in this concept is done using a Master key. Bouganim et al. have presented in [17] a solution named "Client HSM". It consists of integrating the module "HW Security Module" at the level of each DB user. This module holds and protects encryption keys to eliminate their exposure.

Despite the efforts made by researchers to improve the protection of encryption keys at the application level, they still need to be developed and improved. The solutions proposed in the literature are not sufficient as well as each solution has its limits, as we have mentioned in our previous works [3], [5], [8]. In addition, the solutions mostly presented are specific in the case when encryption is done at DBMS.

In this context, the present work aims to propose a new solution to protect encryption keys within applications against internal attacks when adopting encryption at the application level. Our solution is original and provides better protection for keys without any requirement of equipment or material. It is based on the outsourcing of the protection of encryption keys defined on the client side to a DB server. The principle is to transform user's queries holding user's encryption keys to new queries that remotely call functions stored on the DB server in order to convert these keys to real keys for encryption and decryption.

Our paper will be presented as follows: Section II consists of giving general overview of DB encryption at the application level. Section III describes our proposed solution and explains its implementation; we discuss in this section the test results provided by the implementation. Finally, the article ends with a conclusion in Section IV.

II. Database Encryption Preliminaries

A. Database Encryption at the Application Level

Encrypting DB at the application level (client side encryption) means that the process of encrypting/decrypting data is done locally on the application before transmitting data to the DB server. This principle is similar to an externalization of the DB in a cloud storage service [6], [16]. The keys in this concept are managed by the application on the client. It holds all the encryption keys, and there is no transmission of data or keys in the clear to the DB server [4] (see Fig. 1).

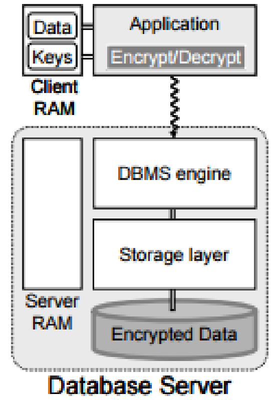


Fig. 1. Database Encryption at Application Level [17].

DB encryption at the application level has many advantages. It resists against attacks coming from administrators. However, it has several disadvantages such as, data and keys are necessarily stored in clear mode on the client, which makes a problematic trust face to the DB users, or even the applications administrator. Both of them cannot always be trusted sites [4], [16].

B. Internal Attack on Encryption at the Application Level

As it has been highlighted by the authors of [4], [5], [8], [16], the DB legitimate operators (including operators that manage applications connected to DB) are considered the main threat compromising the encryption keys when DB is encrypted at the application level. The values of keys and their locations are the vulnerability of this concept. In order to clarify the concept of internal attack when encrypting DB at the application level, we have chosen the attack on DB in both the 2-tier and 3-tier architectures.

In 2-tier client/server architectures, we have 2 essential components [26] including: the client machine (first level) and the DB server (second level). In this client-server model, the DB is encrypted at the client level. The client program accesses DB directly and the encryption keys are in clear on the clients, either in the codes or in the application's configuration files. Thus, they are totally submitted to each user that executes the programs. The users can easily get the encryption keys and attack the DB.

In the 3-tier client-server architectures, we have 3 essential components [26]: the client machine (first level), the application server (second level), and the DB server (third level). In this model, often the keys are stored in clear on the application server and protected using a password in a well-defined server location. Likewise, they are completely submitted to the application server administrator. A simple attack scenario in the 3-tier architecture is a conspiracy attack conducted by a user and the application server administrator. A malicious administrator can reveal all the encryption keys to another user in order to attack DB and decrypt its sensitive data completely without leaving any traces.

III. Proposed Solution

The principle of the proposed solution is to outsource the client's encryption key protection to the DB server. In fact, before sending the user's query, which contains an encryption key (or multiple encryption keys) to the DB server, we transform it into a new query that converts that key (those keys) to the real encryption/decryption key of the DB. For this purpose, we propose and implement two types of solutions. The first one concerns the protection of encryption keys when inserting data into a single sensitive column, while the second concerns the protection of keys when inserting data into multiple sensitive columns.

Let’s consider Table I which has the structure (A) and let us suppose the attribute "Salary" as a sensitive column to be encrypted during insertion and decrypted during the consultation of Table I.

Table I (code, Last_name, First_name, Salary) (A)

A. *Solution1: Inserting in a Single Sensitive Column*

Let’s consider a user who executes an insert query having the following form:

Insert into Tab1 values ('0001', 'elbouchti', 'karim', AES192 ('Mycolor012345678has@ml@p', 2000)); (B)

The user inserts in Tab1 a data line, the column "Salary" will be encrypted with the algorithm AES192 by using the key K="Mycolor012345678has@ml@p". Before sending this query to the DB server, it will be transformed to another query having the syntax below:

Insert into Table I values ('0001','elbouchti','karim', AES192(Func ('Mycolor012345678has@ml@p'), 2000)); (C)

Func() is a function called when the user executes an insert query on a DB table, which has a single sensitive column. The description of Func() is not defined in the files of the user's application or the application server. Its main role is to transform the encryption key value defined in the query that comes out of the user's application into a new encryption key whose value is the real encryption/decryption key of the column (see Fig. 2).

We propose to implement the function Func() in the DB server, it will be one of the objects created inside it. Its model follows the algorithm below:

Func (X)

{K = E (X, H (Nom_table));

Retour (K);}

With:

- E: A symmetric encryption algorithm.
- H: A hash function.
- X: The key defined in the insert statement (user's application files).
- K: the real key to encrypt/decrypt the data of the column.
- Table_name: The table name defined in the insert statement.

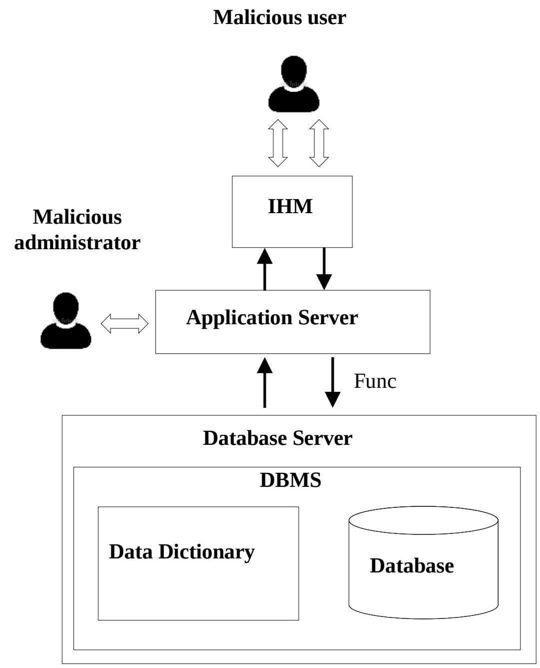


Fig. 2. The Call of Func() Function.

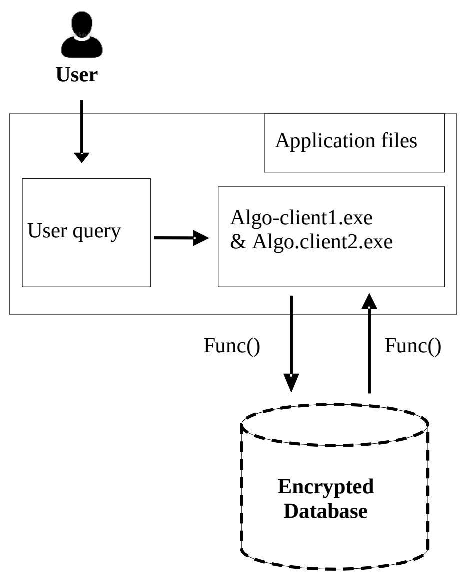


Fig. 3. Process of the Query Converting.

The conversion of query (B) into the query (C) is performed by a transformation algorithm (Algorithm 1 and Algorithm 2 defined below in the two implementations of our solution). These algorithms must appear in executable mode among the files of the application as shown in Fig. 3.

B. *Implementation of the Solution 1*

The implementation of solution 1 has been performed as follows:

- We have developed an application and connect it to a DB named "DB_karim" developed under ORACLE10G.

- We have chosen the AES256 as an encryption algorithm and the MD5 as a hash function to develop the function Func().

When a user executes an insert query, it will be transformed by the algorithm "Algo-client1" into a new query that calls the function Func() of the DB. The code of the "Algo-client1" is defined as follows:

Algorithm 1: Query transformation in solution1.

```
Algo-client1
Input: executed_query
Output: new_query
Begin
Decompose (executed_query)
    Func(key)
    Generate (new_query)
    Execute (new_query)
End
```

The Procedure Decompose() allows the decomposition of the user's query and search exactly the key to substitute by the real encryption key.

The function Func() takes the encryption key defined in the user's query and delivers the real encryption key.

The Function Generate() reformulates the new query with the real encryption key and subsequently send and execute this query using Execute().

The test of "Algo-client1" was performed on the "agent" table which has the following structure:

agent (code, first_name, last_name, dose)   (D)

In this table, we have defined the column "dose" as a sensitive column. It will be protected by an encryption using the AES256 algorithm and the key k = '@mysonmyson@123'. Table I and Table II show the results obtained before and after the encryption

C. Solution 2: Insert in Multiple Sensitive Columns

The principle of protecting encryption keys in a query when inserting data in multiple sensitive columns is similar to solution 1. The main goal is to protect the encryption key of each column distinctly from the other column keys.

Let's consider a user that executes an insert query having the following form:

Insert into Table I values ('0001', AES192('Mycolor442266775hrt@HH@T','elbouchti'), AES192('Mycolormybeautyy@nn@x','karim'),AES192('Mycolor012345678has@ml@p', 2000);   (E)

Before sending that query, it will be transformed into the query (F) having the syntax:

Insert into Table I values ('0001', AES192(Func_Mc('Mycolor442266775hrt@ HH @ T','elbouchti')), AES192(Func_Mc('Mycolormybeautyy@ nn @ x','karim')), AES192(Func_Mc('Mycolor012345678has @ ml @ p')), 2000);   (F)

TABLE. I. THE TABLE "AGENT" BEFORE ENCRYPTION

code	first-name	last_name	dose
1000	Karim	El bouchti	10
1001	hamid	Afane	10
1002	adil	Faris	15
1003	amina	Lemnawar	04
1004	tayebi	El bouchti	06
1005	amina	Aghbal	09

TABLE. II. THE TABLE "AGENT" AFTER ENCRYPTION

code	first-name	last_name	dose
1000	Karim	El bouchti	3C0354B2295D6898C 5BB2731CB6ADCB9
1001	hamid	afane	3C0354B2295D6898C 5BB2731CB6ADCB9
1002	adil	faris	ADE96B60B117A764 F5870665DF4F3D7D
1003	amina	lemnawar	88659CA5130C28136 F6AD7B37F74E531
1004	tayebi	Elbouchti	29152E11B0AEA7CD 69F4327E6AD4730D
1005	amina	aghbal	C1EDB798F41A5C19 F18B2A3E35D6ED92

In this case, the Func_Mc() function is defined and stored in the DB server. It substitutes the keys defined in the query (E) that comes out from the user's application into the real keys of encryption/decryption.We suggest implementing Func_Mc() in the DB server among its objects. It should follow the algorithm below:

Func_Mc (X)

{K = E (X, H ((Nom_table) || (Nom_col)) || (Nom_DB) || Sum (id_col, Id_tab)));

Retour (K) ;}

With:

- E: A symmetric encryption algorithm.
- H: A hash function.
- X: The key defined in the insert statement (user's application files)
- K: the real key to encrypt/decrypt the data of the column.
- Nom_col: The column name defined in the insert statement.
- Nom_DB: The name of DB.
- Sum (id_col, Id_tab): the sum of the column and table identifiers.

D. Implementation of the Solution 2

The implementation of the solution 2 was performed using the same tools deployed in the solution1. We have chosen the AES256 as an encryption algorithm and the MD5 as a hash function to develop the function Func_Mc().

The algorithm "Algo-client2" presented below, transforms the user's query into a new one which calls Func_Mc() before its execution on the DB server.

Algorithm 2: Query transformation in solution 2.

Algo-client2
Input: executed_query Output: new_query Begin Decompose (executed_query) Func_Mc(key) Generate (new_query) Execute (new_query) End

We have tested the algorithm "Algo-client2" on the table "travailleur" which has the structure (G) below:

travailleur (Matricule, dose_prof, dose_sup, dose_neut, dose_int) (G)

We have considered all the "travailleur" columns as sensitive columns. Thus, these columns will be protected by an encryption according to the elements of the Table III and Table IV:

Table V and Table VI show the results obtained when inserting data in "travailleur" table before and after the encryption.

E. Results and Discussions

The results obtained when implementing our two solutions allow us to deduce that both solutions provide better protection of encryption keys, no malicious act performed by the client to hold encryption keys can succeed. The real values of keys are not defined on the client.

Outsourcing the protection of the encryption keys defined on the client side to a DB server is an effective way to secure keys, especially if the server is a trusted site. Our solutions are well adapted to this concept; they are conditioned by a high level of trust at DB server and its operators. Shmueli et al. have mentioned in [11] that the level of trust in the DB server is a fundamental criterion of a DB encryption solution. The partial trust scenario is one of the three levels of trust mentioned in their manuscript. It means that the DB server, together with its memory and the DBMS software are trusted, but the secondary storage that it uses is not. Thus, we choose to integrate our solution into this concept.

TABLE. III. THE ENCRYPTION KEYS ASSOCIATED TO THE "TRAVAILLEUR" COLUMNS

Sensitive column	Algorithm	Encryption key defined in the user query
Matricule	AES256	&@I@encrypt@my@
dose_prof	AES256	&@I@encrypt@he@
dose_sup	AES256	&@I@encrypt@sh@
dose_neut	AES256	&@I@encrypt@it@
dose_int	AES256	&@I@encrypt@we@

TABLE. IV. THE ENCRYPTION KEY GENERATED BY FUNC_MC()

Sensitive column	Encryption key defined in the user query	The encryption key generated by Func_Mc()
Matricule	&@I@encrypt@my@	B50C849772BA517C 27C52327D2CF06B9
dose_prof	&@I@encrypt@he@	E03D47BBED9EA464 AB0BD0DD6B9DF1DA
dose_sup	&@I@encrypt@sh@	F2AB5228D454FF263 044C4D43D92AD50
dose_neut	&@I@encrypt@it@	16307E27C51A178C B731B6F8BDC29CC8
dose_int	&@I@encrypt@we@	E817629DA45424B34 CFCDE4B4B79ADC2

TABLE. V. THE TABLE "TRAVAILLEUR" BEFORE ENCRYPTION

Matricule	dose_prof	dose_sup	dose_neut	dose_int
A00001	1,2	10,2	1,2	0
A00002	2,4	5,47	0,2	0
A00003	6,5	7,21	0,8	0
A00004	2,8	3,55	0	0
A00005	1,1	4,7	0,22	0
A00006	6,5	8,6	0,14	0

TABLE. VI. THE TABLE "TRAVAILLEUR" AFTER ENCRYPTION

Matricule	dose_prof	dose_sup	dose_neut	dose_int
372E7E32A13F7DB9F7FCB 9FBECAEBD3D	FAA8DB86C358211C261058 A7AEB13CC4	48F105C92DBA3DDC5913 C82A9B810BF0	0547D9BDA3BA07B0C1C9 9EC1805247C1	8B954F14B94DFEBB85FE 222B16AE527F
767D2CF95410ECFB0F34C EC0A1B6F466	29C9A2C18ABD74B321FB59 523B85CFFE	0C480ADBEF9CCB14C439 77C5174F035C	F41E9DED97E2112BE602E B10E6888A09	8B954F14B94DFEBB85FE 222B16AE527F
62709892E338E435543A204 719A83198	BB5FB6DFCDD9E1EAC436 ABCA446A2DDA	F55467102D30CDD54E24F D5D18A4453B	BBD0714D33B5DD58E39D 952E338E8BBB	8B954F14B94DFEBB85FE 222B16AE527F
04EA3E21C176BFCD71EE3 BA7DDB8CBA3	786033AD9863E08F31D8936 E5A06E716	5B2FCAC0F7418386B5AF B265002D09E9	94AA0C35085D49DFD6CC 2F4CB5606CDF	8B954F14B94DFEBB85FE 222B16AE527F
4197B8E915413B5F6BD3B1 D93950731C	3FA4CA64C8873E2E01A0C4 3D242403D6	6DF8E9AB7A60CA0B85B1 FDFEBF1B2343	7CF257E8293174EF6980A6 7E9E80645E	8B954F14B94DFEBB85FE 222B16AE527F
586F84DB7D48EB80B2BFB 156528030C4	BB5FB6DFCDD9E1EAC436 ABCA446A2DDA	DAA03A0B62D6A9B9C902 788A9515A2D6	F575A63E49A188328CAB3 2A1AF9A342D	8B954F14B94DFEBB85FE 222B16AE527F

The proposed solutions are well adapted for 2 and 3 tier architectures. For each architecture, the key protection functions (Func () and Func_Mc ()) are implemented inside the DB server. When we choose to manage the DB inside a Cloud (DBAAS), the functions (Func () and Func_Mc ()) will be included among the outsourced DB objects.

Otherwise, these solutions can be compromised by a specific attack, which remains the retro-engineering attack. An attacker can reveal the source code from the executables of the solution files (Algo-client1 and Algo-client2) by using special tools. The attacker can get the return of the functions and can determine how these functions provide the new keys. In this case, we suggest putting all the code of the user's queries in stored procedures, which will also be placed on the DB server.

IV. Conclusion

Databases are the favorite targets of attackers due to the values of the information they contain and their volume. They are vulnerable to several typologies of attacks, especially the internal attacks. When implementing an encryption strategy at the application level to secure DB, the internal attack threats become more dangerous for the sensitive DB, and the probability of attacking the encryption keys is considerable. The attacker can use them to decrypt all sensitive data without leaving any traces

In this work, we have proposed and implemented two solutions that protect encryption keys inside applications. The principle of these solutions is to transform the encryption keys defined at the user's query to other keys considered the real keys for encryption and decryption. This transformation is performed using functions stored within DB server. Our solution is considered efficient and simple when implementation, and it is more adapted to a trusted DB server. In forthcoming works, we aim to adapt our solutions to cover more complicated attacks made by the internal attackers on sensitive DB.

References

[1] I. Homoliak, J.Guarnizo, Y.Elovici, M.Ochoa, “Insight into insiders and it: A survey of insider threat taxonomies, analysis, modeling, and countermeasures” ACM Computing Surveys, New York, NY, USA, vol.52, no.2, pp.30, April 2019.

[2] M. Zabihimayvan, D. Doran, “Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection“, IEEE International C onference on Fuzzy Systems,2019, in press.

[3] K. El Bouchti, S. Ziti, F. Omary, N. Kharmoum, “A New Database Encryption Model Based on Encryption Classes” Journal of Computer Science, vol.15,no.6, pp.844-854, June 2019.

[4] K. El bouchti, N. Kharmoum, S. Ziti, F. Omary, “A new approach to prevent internal attacks on Database encryption keys” Proceedings of the International Conference Scientific Days Applied Sciences, Larache, Morocco,pp.15-16, February 2019.

[5] K. El bouchti, S.ZITI, F.OMARY, “A new approach to protect encryption keys in Database Management System”, Proceedings of the International Conference Modern Intelligent Systems Concepts, Rabat, Morocco, pp.12-13, December 2018.

[6] C.Priebe, K.Vaswani, M.Costa,M, “Enclavedb: A secure database using sgx“, IEEE Symposium on Security and Privacy (SP), USA, pp.264-278, May 2018.

[7] Y.Elovici, R.Vaisenberg, E.Shmueli. (2018). U.S. Patent No. 9,934,388. Washington, DC: U.S. Patent and Trademark Office.

[8] K.El bouchti, S.Ziti, Y.Ghazali, N.Kharmoum, “Sécurité des Bases de Données : Menaces principales et solutions de chiffrement existantes“, Proceedings of the JDSIRT Conference Information Systems, Networks Telecommunications, Meckness, Morocco, pp.13 ,November 2018.

[9] Hashim, Hassan B, “Challenges and Security Vulnerabilities to Impact on Database Systems“, Al-Mustansiriyah Journal of Science,vol. 29, no.2, pp.117-125, 2018.

[10] Zainab.S.Alwan, Manal.F.Younis, “Detection and Prevention of SQL Injection Attack: A Survey“, International Journal of Computer Science and Mobile Computing, vol. 6, no 8, pp. 5-17, 2017.

[11] E.Shmueli, R.Vaisenberg, E.Gudes, Y. Elovici, “Implementing a database encryption solution, design and implementation issues“, Computers & security, vol.44, pp.33-50, 2014.

[12] B.H.Chen, P.Y.Cheung, P.Y.Cheung, Y.K.Kwok, “Cypherdb: A novel architecture for outsourcing secure database processing“, IEEE Transactions on Cloud Computing, vol.6, no.2, pp.372-386, 2018.

[13] Deepicata. N. Soni, “Database Security: Threats and Security Techniques”, International Journal of Advanced Research in Computer Science and Software Engineering, vol.5, no.5, pp.621-624, 2015.

[14] J.D. Bokefode, S.A. Ubale, S.S. Apte, D.G. Modani, “Analysis of dac mac rbac access control based models for security. Analysis“,International Journal of Computer Applications,vol. 104, no 5, pp.6-13, 2014.

[15] I.Basharat, F.Azam, A. Muzaffar,”Database Security and Encryption: A Survey Study”, International Journal of Computer Applications, vol. 47,no.12,pp.28-34, June 2012.

[16] S. Jacob, “Protection cryptographique des bases de données: conception et cryptanalyse,“ Ph.DThesis, Université Pierre et Marie Curie-Paris VI, 2012.

[17] L.Bouganim, Y.Guo, “Database encryption”, In Encyclopedia of Cryptography and Security, Springer US, pp. 307-312, 2012.

[18] E.Shmueli, R.Vaisenberg,Y.Elovici, C.Glezer, “ Database encryption: an overview of contemporary challenges and design considerations,” ACM SIGMOD Record, vol.38, no.3,pp. 29-34, 2010.

[19] Y.Ding, K.Klein, “Model-driven application-level encryption for the privacy of e-health data”, International Conference in Availability, Reliability, and Security, IEEE, Krakov, Poland, pp. 341-346, February 2010.

[20] L.Liu, J.Gai, “A new lightweight database encryption scheme transparent to applications”, 6th IEEE International Conference on Industrial Informatics, Daejeon, South Korea, pp.135-140, July 2008.

[21] S.Sesay, Z. Yang, J.Chen, D. Xu, “A secure database encryption scheme”, Procceding of the Consumer Communications and Networking Conference, Second IEEE, Las Vegas, USA, pp.49-53, January 2005.

[22] Mattsson, U. T, “ A practical implementation of transparent encryption and separation of duties in enterprise databases: protection against external and internal attacks on databases”, Proceeding of the International Conference on E-Commerce Technology, Munich, Germany, pp. 559-565, July 2005.

[23] Y.Elovici,R.Waisenberg, E.Shmueli, E.Gudes, “A structure preserving database encryption scheme”, In Workshop on Secure Data Management, Berlin, Germany, pp.28-40, 2004.

[24] J.Wang, T.Wang, Z.Yang, Y.Mao, N.Mi, & B.Sheng, “Seina: A stealthy and effective internal attack in hadoop systems”, In International Conference on Computing, Networking and Communications, IEEE, Santa Clara, USA, pp.525-530, January 2017.

[25] T.Wang, N.S.Nguyen, N, J.Wang, T.Li, X.Zhang, N.Mi, & B.Sheng, “Rover: Robust and verifiable erasure code for hadoop distributed file systems”, The 27th International Conference on Computer Communication and Networks, Hangzhou, China, pp.1-9, July 2018.

[26] M.Kambalyal, 3-tier architecture. Retrieved On, vol.2, p.34.

Neural Network Supported Chemiresistor Array System for Detection of NO_2 Gas Pollution in Smart Cities (NN-CAS)

Mahmoud Zaki Iskandarani
Faculty of Engineering
Al-Ahliyya Amman University
Amman, Jordan

***Abstract*—Neural Networks supported Chemiresistor array system is designed and laboratory tested for the detection of emissive gasses from vehicles and other sources of pollution. The designed and tested system is based on an integrated PbPc array of chemiresistors that sends signals corresponding to emitted NO_2 gas to Signal Processing Unit. The process comprises using relative conductivity values of Edge sensors to Central sensor for detected gas as an indicator of response characteristics and profiling for NO_2 gas pollution level. The process continues up to the limit where Edge Sensor values for relative conductivity equates, then the relative conductivity for the Edge Sensors is used as a control value to shut down the sampling system and send a warning message of excessive pollution. Pollution could be due to a number of factors besides vehicles, such as gas leaks. Optimization of array elements response is carried out using Neural Networks (Back Propagation Algorithm). The proposed system is promising and could further be developed to become a vital and integrated part of Intelligent Transportation Systems (ITS) in order to monitor emission of hazardous gases, and could be integrated with Road Side Units (RSUs) of urban areas in smart cities.**

Keywords—Gases; chemiresistors; neural networks; sensor array; correlation; road side unit; intelligent transportation systems; smart cities

I. Introduction

Emissions of NOx and NO2 from vehicles are critical to quality of air particularly in urban areas, and could very well affects air quality at regional and global levels.

Recently, two important factors are considered that contributes to pollution and concentration of NO_2: NO_x in urban areas:

1) The ratio of NO_x that is NO_2 coming out of vehicles exhausts.

2) Diesel engines emissions of NO_x

Congested cities and their residents exposed to levels of NO_2 gas that often exceed the acceptable air quality standards. Due mainly to diesel cars. The level of contribution of NO_x by the diesel car is determined as per area and number of vehicles and congestion levels. However, it is found that large number of NO_2 parts in the NO_x emissions of diesel engines is mainly a function of intense road traffic usually on artery roads.

NO_x contains NO and NO_2, where NO_2 is critical as it has an adverse health effects in urban areas. Diesel engines are not fitted with efficient systems for removing NO_x emissions similar to petrol engines, thus, resulting in higher ambient concentration of NO_2 in urban and major cities.

The primary health effects attributable to NO_2 are related to respiratory conditions. Inhalation of NO_2 causes inflammation in the lungs, affecting immunity to lung infections and resulting in loss of breath, wheezing, coughing and bronchitis with possibility of developing asthma. NO_2 can cause has both acute and chronic health effects.

Studies showed that Lead Phthalocyanine (PbPc) is very sensitive complex to Dioxide gases; specifically NO_2, where its conductivity affected more by the adsorption of gases as a charge-transfer complex is formed between the Phthalocyanine donor and the gas acceptor.

II. Related Work

Urbanization adds pressure to the resources such as energy, water, sanitation, and public services. Thus, socio-economic and environmental issues have become closely related. Cities contribute to environmental change on local, regional, and global scales. Studies showed that cities accounts for large amount of global greenhouse gas emissions as a function of energy consumption. City planners and researchers worldwide are investigating ways to control traffic in order to improve air quality, and provide enhanced living conditions [1-4].

The solution is in making cities “smarter” through different approaches to resources management and infrastructure, and by concentrating on greener environment, and smart governance, which will result in a better quality of living for citizens. This can be enabled by utilization of Information and Communication Technologies (ICTs) tools, which can provide eco-friendly solutions for cities. Such work lead to the concept of Smart Cities, whereby the vision is to include the basic services in the city, such as clean water, clean environment, energy and infrastructure, for all citizens, which can be achieved by creating smart environment that covers:

1) Environmental Sustainability

2) Energy Consumption Control

This can be accomplished by focusing on smart transportation that connects different modes of transportation into an integrated system, thus, giving city planners the ability to better control the flow of traffic.

Analysis and modelling of urban air quality was most of the time based on the assumption that vehicles on the road perform similarly to the way they do under development environments, this can lead to inaccurate prediction of the vehicles effect and contribution to air pollution and harmful emissions. The application of PbPc sensor array chemiresistors to the detection and subsequent analysis of gases emitted round urban areas and congested cities such as NO_2, should provide reliable metric that can also be used in both real life and in development environments and can help in narrowing the gap between the development sites and real life applications in real time [5-9].

As Smart Cities are associated with a higher quality of life, technology makes it possible to compile massive amounts of real-time data to optimize the urban infrastructure, thereby improving the efficiency of public and transport services.

A chemiresistor array system (CAS) is generally recognized as a system that encompasses array of chemical sensors with selective detection capabilities and pattern recognition capability, able to specify individual vapor components or combination of vapors. The CAS recognizes the presence of a chemical through fingerprinting of its chemical elements using an array of sensors backed by intelligent software for pattern recognition [10-15].

There are two major components forming the CAS:

1) Chemiresistor Sensing Array (CSA).

2) Intelligent Part employing Artificial Neural Networks (ANN).

Such a combination makes CAS a promising tool for detection of chemicals and hazardous gases. Each chemical produces a unique characteristic of its own, once exposed to the chemiresistor sensing array. The experimental data is used to train an intelligent classification system, such as Neural Networks in order to optimize the CAS characterisitics and to provide an ability to predict future values based on chemical level changes [16-20].

CAS detects chemicals by interacting with its CSA responsive materials, resulting in a change in the material characteristics and producing a unique response associated with a specific chemical or gas.

ANN is a learning and classification algorithm, and can also be used as an optimizing algorithm. ANN changes its input, hidden, and output neuron weights to interrelate and correlate complex relationships among input-output variables. Backpropagation (BP) algorithm, is an affective ANN technique, which is an iterative gradient algorithm aims at decreasing the root mean square error.

In this paper, a fresh approach to the use and application of Chemiresistor Arrays System is proposed which utilizes chemical sensing, in particular NO_2 together with Neural Networks optimization. Such approach will support environmental mobility of vehicles through big data collection and analysis and vehicle to infrastructure interface (V2I). The system can be further developed to support traffic light control and green wave for certain vehicles such as diesel engines when integrated with Road Side Units (RSUs) and interfaced using wireless communication systems [21-22].

III. Materials and Methods

Chemiresistor array units are used for the tests. The NO_2 detection system employs a number of chemiresistors with vacuum sublimed PbPc films of uniform thickness on Sapphire (α-Al_2O_3) substrates. Fig. 1 shows 4-electrodes, 3-chemiresistor array device used in the testing, while Fig. 2 shows a cross sectional view of each chemiresistor within the array. Testing of the devices response to donor gases, in particular NO_2 using two devices is carried out as shown in Fig. 3.

Back Propagation Algorithm (BP) is used to carry out training of the Neural Networks system in order to optimize the response characteristics of the used chemiresistor arrays shown in Fig. 4.

Back Propagation (BP) works by repeatedly modifies the weights of the connections in the network in order to minimize the difference between the actual output of the network and the desired output. The internal 'hidden' units which are not part of the input or output stores within their weights the important features of the learnt pattern(s), which are captured by the interactions of these units. The algorithm is used to efficiently train a neural network through a chain rule, where, after each forward pass through a network, backpropagation performs a backward pass while adjusting the network weights and biases.

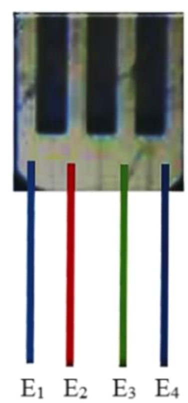


Fig. 1. 3-Chemiresistor Integrated Array Device.

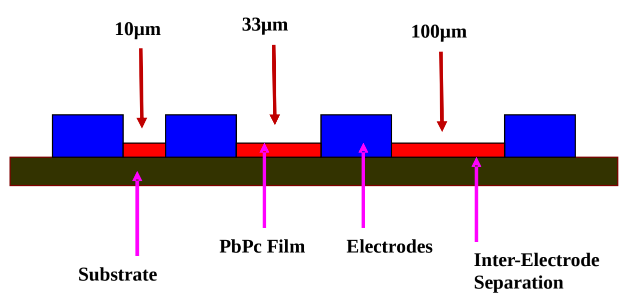


Fig. 2. Cross Section of the used PbPc Chemiresistor Array.

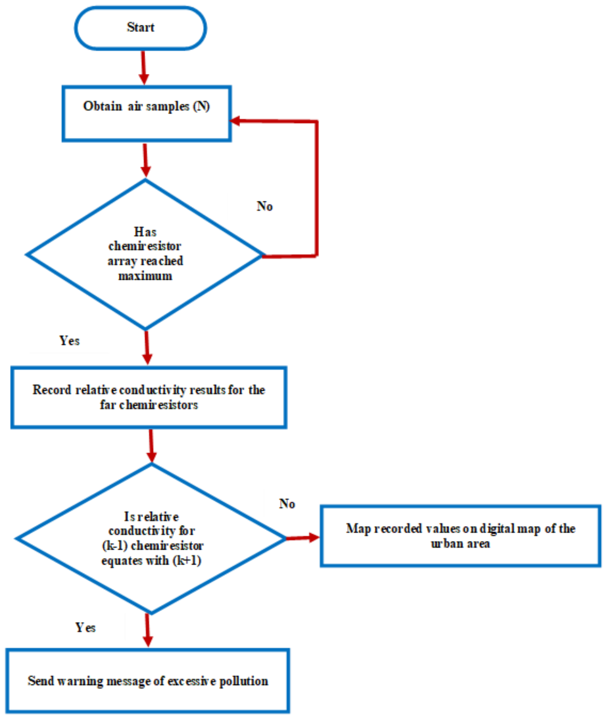


Fig. 3. PbPc Array Algorithm for NO_2 Detection.

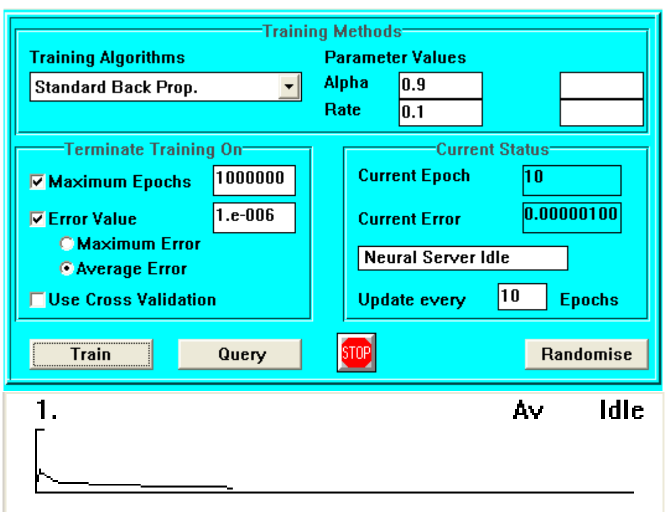


Fig. 4. Neural Networks (BP) Traninig System with Training Curve.

Optimization of the CAS response using BP is carried out by reducing the error function through minimization of the error function (cost function) described by equation (1).

$$E(t) = \frac{1}{2}\left(d_{L(i+1)}(t) - a_{L(i+1)}(t)\right)^2. \quad (1)$$

Where;

$d_i(t)$: The desired response

$a_i(t)$: The actual response

To carry out error mimimization using BP, a gradient descent rule is used to update weights between output and hidden layers and hidden and input layers as shown in equation (2).

$$W_{L(i+1),\,L(i)}(t+\Delta t) = W_{L(i+1),\,L(i)}(t) - Alpha\left(\frac{\partial E(t)}{\partial W_{L(i+1),\,L(i)}}\right). \quad (2)$$

Where;

$Alpha$: Learning Rate.

The learning Rate is increased from 0.1 to 0.9 and decreased from 0.9 to 0.1 as shown in Fig. 4, in order to compute weight updates using equations (3) and (4).

$$Alpha_{increasing} = Alpha_{\min} + \left(Alpha_{\max} - Alpha_{\min}\right)\left(\frac{(Max\ Epochs - Current\ Epoch)}{Max\ Epochs}\right). \quad (3)$$

$$Alpha_{decreasing} = Alpha_{\max} - \left(Alpha_{\max} - Alpha_{\min}\right)\left(\frac{(Max\ Epochs - Current\ Epoch)}{Max\ Epochs}\right). \quad (4)$$

Results

Tables I and II show data for two PbPc sensor arrays used in validating the designed system, while Fig. 5 shows the Neural Network model used for training with distributed weights.

TABLE. I. SENSOR ARRAY 1

Real Test	**Normalized conductivity in relation to Inter-Electrode Separation**		
NO_2 Levels ppm	**Chemi1,2 10:33**	**Chemi1,3 10:100**	**Chemi2,3 33:100**
0	0	0	0
1	0.66	0.34	0.51
3	0.69	0.44	0.64
5	0.71	0.47	0.66
7	0.72	0.48	0.67
9	0.73	0.49	0.73

TABLE. II. SENSOR ARRAY 2

Real Test	**Normalized conductivity in relation to Inter-Electrode Separation**		
NO_2 Levels ppm	**Chemi1,2 10:33**	**Chemi1,3 10:100**	**Chemi2,3 33:100**
0	0	0	0
1	0.64	0.34	0.54
3	0.67	0.44	0.65
5	0.68	0.46	0.68

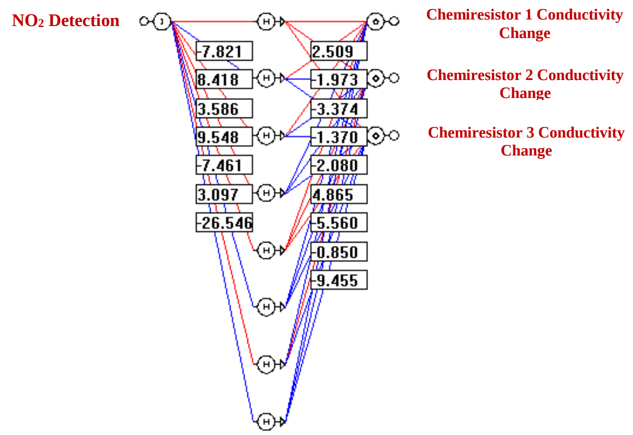


Fig. 5. Neural Networks Model used for Trianing, Optimization, and Prediction.

IV. Discussion and Conclusions

Tables III and IV show the predicted results using Neural Networks (BP) System, while Fig. 6 and 7 show the optimized response curves for the PbPc arrays, with Fig. 8 and 9 showing the convergence factor in relation to the relative conductivity change of the PbPc sensor array as a function of NO_2 gas concentration.

The relative conductivity of the Central Sensor can be approximated and related to the edge sensors, using the expression in equation (5). In addition, the actual convergence factor (k) for each gas concentration is also calculated using the expression in equations (5), together with Neural Networks predicted data, which should fulfill the criteria descripted by the expression in equation (6), while latching conditions for the system of array sensors is applied using the average convergence factor (k_{Avg}) as shown in equation (7).

$$Convergence\,(Central\,Sensor) = \left[\frac{\sum Relative\,Conductivity\,(Edge\,Sensors)}{Total\,Number\,of\,\,Array\,Sensors}\right]. \tag{5}$$

$$Convergence\,(Array) = \left(\frac{Lim\,(Relative\,Conductivity\,Far\,Left\,Edge\,Sensor)}{Lim\,(Relative\,Conductivity\,Far\,Right\,Edge\,Sensor)}\right) = 1.. \tag{6}$$

$$Latching = \left[\frac{\sum Relative\,Conductivity\,(Edge\,Sensors)}{Average\,Convergence\,Factor\,(k_{Avg})}\right].. \tag{7}$$

The results from the two PbPc sensor array devices showed different response sensitivities towards NO_2 gas, whereby array 1 (average convergence factor 3.05) latched at higher concentration levels compared with array 2 (average convergence factor 3.3). Both devices have comparable results up to 5 PPM of NO_2 concentration. This is a design issue, which necessitates the use of Neural Networks to predict data to enable design optimization and performance enhancement.

Fig. 6 and 7 show relative conductivity response of both devices, which presents the power increase of relative conductiviy of the Edge and Center elements of the array. They also show the convergence process between the two Edge elements as they equate to same value. This is also a design issue as the electrode separation between each Edge element and the Center element is approximately a factor of 3.

Fig. 8 and 9 present a clearer view of the field interaction between array elements in the form of the convergence factor as it initially varies before it decreases and converges to the value of approximately 3. Thus, conforms to the electrode separation in the original design.

TABLE. III. Neural Networks Optimization for Array 1

Prediction	Normalized conductivity in relation to Inter-Electrode Separation			Convergence Factor
NO_2 Levels PPM	Chemi1,2 10:33	Chemi1,3 10:100	Chemi2,3 33:100	k
0.000	0.000	0.000	0.000	0.0
0.004	0.002	0.001	0.001	3.0
0.008	0.004	0.002	0.003	3.5
0.010	0.005	0.002	0.004	4.5
0.040	0.022	0.010	0.016	3.8
0.080	0.050	0.022	0.034	3.8
0.100	0.065	0.030	0.044	3.6
0.400	0.375	0.157	0.240	3.9
0.800	0.618	0.300	0.453	3.6
1.000	0.660	0.340	0.510	3.4
1.400	0.690	0.383	0.570	3.3
1.800	0.696	0.404	0.600	3.2
2.000	0.695	0.412	0.608	3.2
2.400	0.692	0.425	0.623	3.1
2.800	0.690	0.435	0.635	3.0
3.000	0.690	0.440	0.640	3.0
3.400	0.691	0.449	0.648	3.0
3.800	0.695	0.456	0.654	2.9
4.000	0.697	0.459	0.656	2.9
4.400	0.703	0.463	0.660	2.9
4.800	0.710	0.470	0.660	2.9
5.000	0.710	0.470	0.660	2.9
5.400	0.714	0.473	0.660	2.9
5.800	0.716	0.475	0.661	2.9
6.000	0.717	0.476	0.661	2.9
6.400	0.719	0.477	0.663	2.9
6.800	0.720	0.479	0.667	2.9
7.000	0.720	0.480	0.670	2.9
7.400	0.721	0.482	0.677	2.9
7.800	0.722	0.483	0.687	2.9
8.000	0.723	0.484	0.697	2.9
8.400	0.725	0.486	0.703	2.9
8.800	0.728	0.489	0.722	3
9.000	0.730	0.490	0.730	3

TABLE. IV. NEURAL NETWORKS OPTIMIZATION FOR ARRAY 2

Prediction	**Normalized conductivity in relation to Inter-Electrode Separation**			**Convergence Factor**
NO_2 Levels ppm	**Chemi1,2 10:33**	**Chemi1,3 10:100**	**Chemi2,3 33:100**	**k**
0.000	0.000	0.000	0.000	0
0.004	0.002	0.001	0.001	3
0.008	0.004	0.002	0.003	3.5
0.010	0.005	0.002	0.004	4.5
0.040	0.021	0.010	0.016	3.7
0.080	0.046	0.022	0.035	3.7
0.100	0.060	0.028	0.046	3.8
0.400	0.350	0.152	0.255	4
0.800	0.595	0.298	0.482	3.6
1.000	0.640	0.340	0.540	3.4
1.400	0.673	0.386	0.597	3.3
1.800	0.679	0.408	0.621	3.2
2.000	0.678	0.416	0.628	3.1
2.400	0.674	0.428	0.638	3.1
2.800	0.671	0.436	0.646	3
3.000	0.670	0.440	0.650	3
3.400	0.670	0.446	0.656	3
3.800	0.671	0.451	0.662	2.95
4.000	0.673	0.453	0.665	2.95
4.400	0.676	0.456	0.671	2.95
4.800	0.679	0.459	0.677	2.95
5.000	0.680	0.460	0.680	2.96

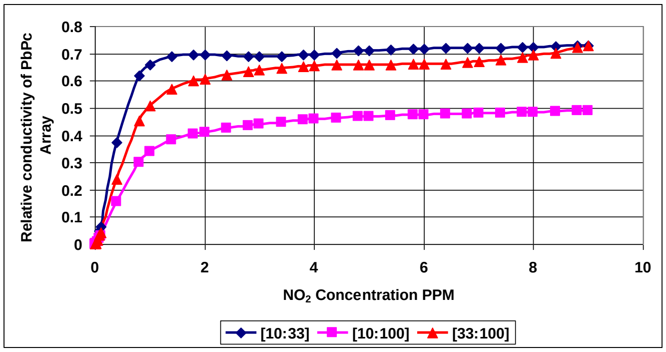


Fig. 6. Predicted and optimized sensor response for array 1

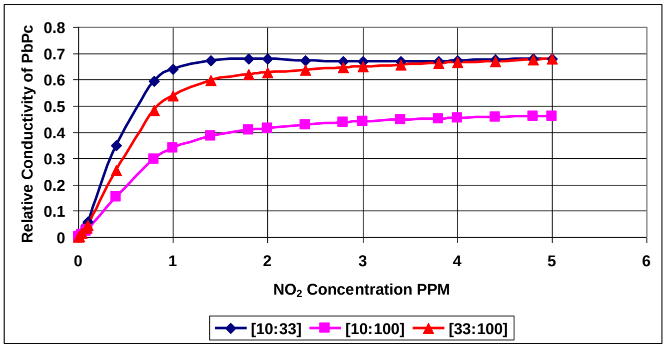


Fig. 7. Predicted and Optimized Sensor Response for Array 2.

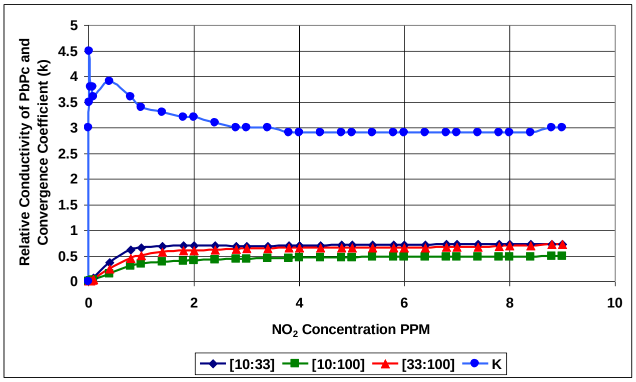


Fig. 8. Predicted Sensor Response and Convergence Coefficient for Array 1.

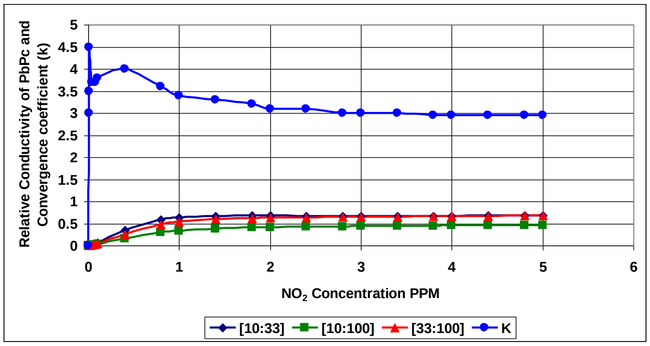


Fig. 9. Predicted Sensor Response and Convergence Coefficient for Array 2.

Employing Neural Networks allowed for:

1) Acquiring results for other NO_2 concentration levels.

2) Design optimization based on prediction.

3) Detection of expected increase or decrease of NO_2 in an area as a function of current values correlated with traffic volume and other sources of NO2 emission.

In conclusion, the design and testing of the PbPc sensor arrays was successful and more so with the incorporation of Neural Networks. Smart cities and smart transportation systems, aim to provide less polluted urban areas and such chemiresistor arrays can be very useful in this context. Developing A wireless Sensor Networks (WSN) version of the PbPc array will certainly advance its application and enhance the monitoring and reporting facilities through wireless data routing and collection to a sink with a cloud interface to control centers.

REFERENCES

[1] D. Carslawa, G. Rhys-Tyler, "New insights from comprehensive on-road measurements of NO_x, NO_2 and NH_3 from vehicle emission remote sensing in London, UK," Atmospheric Environment, vol. 81, pp. 339–347, 2013.

[2] B. Degraeuwe, P. Thunis, A. Clappier, M. Weiss, W. Lefebvre, S. Janssen , S. Vranckx, "Impact of passenger car NO_x emissions and NO_2 fractions on urban NO_2 pollution e Scenario analysis for the city of Antwerp, Belgium," Atmospheric Environment, vol. 126, pp. 218–224, 2016.

[3] B. Degraeuwea, P. Thunisa, A. Clappierb, M. Weissc, W. Lefebvred, S. Janssend, S. Vranckxd, "Impact of passenger car NO_X emissions on urban NO2 pollution – Scenario analysis for 8 European cities," Atmospheric Environment, vol. 171, pp. 330–337, 2017.

[4] J. Jonson, J. Borken-Kleefeld, D. Simpson, A. Ny´ıri1 , M. Posch, C. Heyes, "Impact of excess NO_x emissions from diesel cars on air quality, public health and eutrophication in Europe," Environmental Research Letters, vol. 12, pp. 1–11, 2017.

[5] R. Smit, P. Kingston, "Measuring On-Road Vehicle Emissions with Multiple Instruments Including Remote Sensing," Atmoshphere, vol. 10, pp. 1–17, 2019.

[6] M. Otto, P. Ramacher, M. Karl, J. Bieser, J. Jalkanen, L. Johansson, "Urban population exposure to NOx emissions from local shipping in three Baltic Sea harbour cities – a generic approach," atmospheric Chemistry and Physiscs, vol. 19, pp. 9153–9179, 2019.

[7] N. Zulauf , J. Dröge, D. Klingelhöfer, M. Braun , G. Oremek, D. Groneberg, "Indoor Air Pollution in Cars: An Update on Novel Insights," International Journal of Environmental Research and Public Health, vol. 16, pp. 1–11, 2019.

[8] L. Chatzidiakou, A. Krause1, O. Popoola1 , A. Antonio , M. Kellaway , Y. Han, F. Squires, T. Wang, H. Zhang. Q. Wang, Y. Fan, S. Chen, M. Hu, J. Quint, B. Barratt, F. Kelly, T. Zhu, R. Jones, "Characterising low-cost sensors in highly portable platforms to quantify personal exposure in diverse environments," Atmospheric Measurements Techniques, vol. 12, pp. 4643–4657, 2019.

[9] M. Hasan, V. Chourasia, S. Asutkar, "Breathe-Safe: An IoT based predictive tool for health care"International Journal of Scientific & Technology Research, vol. 8, pp. 2945–2957, 2019.

[10] S. Vishesh, M. Srinath, K. Gubbi, H. Shivu, N. Prashanta, "Portable Low Cost Electronic Nose for Instant and Wireless Monitoring of Emission Levels of Vehicles Using Android Mobile Application," IJARCCE, vol. 5, no. 9, pp. 134–140, 2016.

[11] A. Guentner, V. Koren, K. Chikkadi, M. Righettoni, S. Pratsinis, "E-nose sensing of low-ppb formaldehyde in gas mixtures at high relative humidity for breath screening of lung cancer?" ACS Sensors, vol. 1,no. 5, pp. 528–535, 2016.

[12] A. Gongora, J. Monroy, J. Gonzalez-Jimenez, "An Electronic Architecture for Multipurpose Artificial Noses," Journal of Sensors, vol. 2018, pp. 1-9, 2018.

[13] Y. Sun, D. Luo, H. Li, C. Zhu, O. Xu, H. Hosseini, "Detecting and Identifying Industrial Gases by a Method Based on Olfactory Machine at Different Concentrations," Journal of Electrical and Computer Engineering, vol. 2018, Article ID 1092718, pp. 1–9, 2018.

[14] A. Tiele, S. Esfahani, J. Covington, "Design and Development of a Low-Cost, Portable Monitoring Device for Indoor Environment Quality," Journal of Sensors, vol. 2018, Article ID 5353816, pp. 1–14, 2018.

[15] A. Iosifidis, A. Tefas, and I. Pitas, "Approximate kernel extreme learning machine for large scale data classification," Neurocomputing, vol. 219, pp. 210–220, 2017.

[16] Z. Ma, G. Luo, K. Qin, N. Wang, W. Niu, "Weighted Domain Transfer Extreme Learning Machine and Its Online Version for Gas Sensor Drift Compensation in E-Nose Systems," Wireless Communications and Mobile Computing, vol. 2018, Article ID 2308237, pp. 1–17, 2018.

[17] A. Di Gilio, J. Palmisani, G. de Gennaro," An Innovative Methodological Approach for Monitoring and Chemical Characterization of Odors around Industrial Sites," Advances in Meteorology, vol. 2018, Article ID 1567146, pp. 1–8, 2018.

[18] A. Hassan, A. Bermak, "Biologically Inspired Feature Rank Codes for Hardware Friendly Gas Identification with the Array of Gas Sensors," IEEE Sens. J, vol. 16, pp. 5776–5784, 2016.

[19] Y. Wu, T. Liu, S. Ling, J. Szymanski, W. Zhang, S. Su, "Air Quality Monitoring for Vulnerable Groups in Residential Environments Using a Multiple Hazard Gas Detector," Sensors, vol. 19, no. 362, pp. 1–16, 2019.

[20] C. Deng, K. Lv, D. Shi, B. Yang, S. Yu, Z. He, J. Yan, "Enhancing the Discrimination Ability of a Gas Sensor Array Based on a Novel Feature Selection and Fusion Framework," Sensors, vol. 18, no. 1909, pp. 1–17, 2018.

[21] J. Guerrero-Ibáñez, S. Zeadally, J. Contreras-Castillo, "Sensor Technologies for Intelligent Transportation Systems," Sensors, vol. 18, no. 1212, pp. 1–24, 2018.

[22] M. Iskandarani, "Two Dimensional Electronic Nose for Vehicular Central Locking System (E-Nose-V)," IJACSA, vol. 10, no. 6, pp. 63-70, 2019.

A Framework for Detecting Botnet Command and Control Communication over an Encrypted Channel

Zahian Ismail[1], Aman Jantan[2], Mohd. Najwadi Yusoff[3]
School of Computer Sciences
Universiti Sains Malaysia
Pulau Pinang, Malaysia

***Abstract*—Botnet employs advanced evasion techniques to avoid detection. One of the Botnet evasion techniques is by hiding their command and control communication over an encrypted channel like SSL and TLS. This paper provides a Botnet Analysis and Detection System (BADS) framework for detecting Botnet. The BADS framework has been used as a guideline to devise the methodology, and we divided this methodology into six phases: i. data collection, customization, and conversion, ii. feature extraction and feature selection, iii. Botnet prediction and classification, iv. Botnet detection, v. attack notification, and vi. testing and evaluation. We tend to use the machine learning algorithm for Botnet prediction and classification. We also found several challenges in implementing this work. This research aims to detect Botnet over an encrypted channel with high accuracy, fast detection time, and provides autonomous management to the network manager.**

Keywords—Botnet; Botnet Analysis and Detection System (BADS); encrypted channel; machine learning; accuracy; autonomous

I. INTRODUCTION

Botnet has become a significant concern in the computer industry. With users engaging in daily life surfing to the Internet, there was a high risk of becoming a victim of a Botnet attack. The botnet has developed many capabilities, but unfortunately, most of those capabilities are used for attack purposes, such as performing a DDoS attack, spamming, malware spreading, and large computer compromising. Launching a massive DDoS attack is one of the main capabilities of Botnet. For instance, a DDoS attack that happened in the year 2000 is one of the notorious DDoS attacks when the cyber-criminal targeting Yahoo!, Fifa.com, Amazon.com, Dell, Inc., E*TRADE, eBay, and CNN. Another main capability of Botnet is spamming. Several authors like Solomon and Evron (2006) [1] highlight Botnet spamming as a significant concern because of the large amount of distribution of spam, which will use many network resources. Their concern was supported by McAfee Avert Labs [2] which stated that more than 70 percent of spam email caused by Botnet.

Other than DDoS and spamming, Botnet vigorously compromised many computers and tried to develop a vast network of the infected machine through its command and control (C&C) communication. Thus the impact of the attack is enormous. Therefore, Botnet is becoming an increasingly widely-used method by cybercriminals for many purposes, such as gaining recognition from other hackers, financial gain, and many other nefarious activities; hence, all of these affect the users in general. The Botnet is also making antivirus tools ineffective, and bots able to modify registry entries, so they remain active even when the infected machine is booted in a safe mode. Some of the Botnet even respond vigorously if they notice there are efforts made trying to detect their presence.

Considering such capabilities deployed in many Botnets, the effects of Botnet attack are so huge. Botnet brings high risk to national security, intimidates the security of many organizations, either public or private entities, especially caused terrible disturbance and high usage of network resources. Cleaning on the detected system will be very difficult because the volume of network traffic created by bots is massive, thus making it impossible to perform an update on an infected machine (Thomas & Jyoti, 2007) [3]. For this reason, even governments have to spend much money to prevent Botnet attacks. CyberSecurity Malaysia [4] provides the statistic of Botnet attacks in Malaysia, as depicted in Fig. 1. Table I shows Botnet attacks in comparison to other attacks in Malaysia for six consecutive years. Botnet employs many evasion techniques to avoid detection and stay in the network. One of the evasion techniques is by manipulating encrypted channel like SSL and TLS to hide their C&C activities. Zhang (2017) [5] refers Botnet that uses encryption evasion techniques as an advanced Botnet. Burghouwt (2015) [6] states the encryption of C&C traffic as the most crucial evasion technique by Botnet. The Botnet dependency to this evasion technique is due to several reasons; for instance, the increasing number of services and applications that use an encrypted channel to secure the communications and contents.

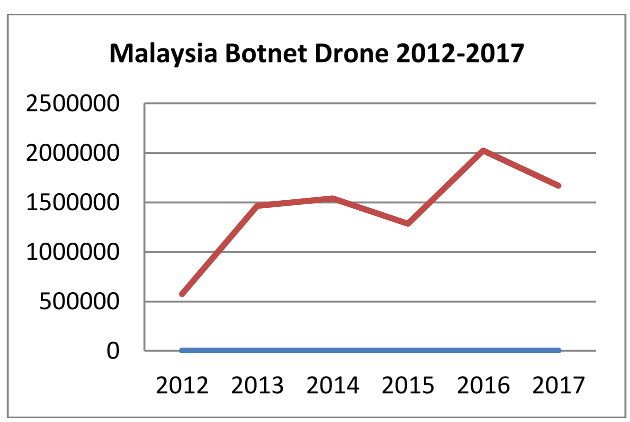


Fig. 1. Malaysia Botnet Statistics 2012-2017 (MyCert, 2017).

TABLE. I. TREATS COMPARISON IN MALAYSIA FROM 2012-2017

Year	DoS	Mal code	Spam	Intrusion	Fraud	Botnet
2012	23	645	526	4326	4001	573401
2013	19	1751	950	2770	4485	1465785
2014	29	716	3650	1125	4477	1539734
2015	38	567	3539	1714	3257	1285605
2016	64	375	518	2328	3612	2026276
2017	40	814	344	2011	3821	1669973

Furthermore, according to Nicholson (2015) [7], social media like Pinterest and Twitter, and email applications use SSL/TLS to encrypt all communications. These social media have many users, thus making them a good target for the attacks because of the potential to compromise the massive host. Moreover, web traffic is allowed almost everywhere. Botnet strains, for example, Storm, Waledac and Rustock, use social media and email applications for their C&C communication to evade detection and stealthily distribute the command and control bots in the network. Hence, Paganini (2014) [8] states that while SSL/TLS encryption improves privacy and integrity, Botnet uses SSL blind spot to avoid detection and leave their C&C covert for as long as possible.

Even though the reports by CyberSecurity Malaysia in Fig. 1 and Table I did not directly state that the Botnet attacks have been performing through Secure Socket Layer (SSL) or any encrypted channel, we consider another report by Gebhart from Electronic Frontier Foundation (EFF) [9] which stated that by 2017 more than 50% of Internet traffic had been protected by HTTPS. The effort of turning the traffic into encrypted one was enthusiastically made since 2010. Therefore, the botmasters are taking advantage of this situation to hide their operation and evade detection. Botnet itself creates a massive impact, and with the implementation of advanced evasion techniques like masquerading in the SSL channel will amplify the impacts. These scenarios have shown the severity of encrypted Botnet attack and therefore become an encouraging factor for developing solutions to detect Botnet over an encrypted channel even though there are other network attacks encrypted in SSL/TLS-enabled protocol.

Additionally, according to Nicholson (2015) [7] and Finley (2017) [10], roughly more than half of all traffic is encrypted mostly by SSL/TLS. By the end of 2016, Gooley (2017) [11] states that 80% of traffic across Google properties was encrypted, and 54% of threats Zscaler blocked are hidden inside SSL traffic. Therefore the use of SSL for the distribution of malicious content is rising too. Recent Botnet strains manipulated this situation and use SSL/TLS channel for their command and control communication. SSL/TLS protects legitimate content but simultaneously provides Botnet with hiding spots, making encrypted channel beneficial for a good guy and bad guy. There are three most active malicious contents referred by Gooley which are Dridex, Vawtrak, and Gootkit; all of them are variants of Botnet, which commonly associated with user credential stealing. Moreover, Rossow & Dietrich (2013) report that detailed C&C traffic analysis shows at least ten prevalent malware families avoided well-known C&C carrier and preferred encrypted channel, and one of them is Zeus. Consequently, the rising of SSL-encrypted traffic increases the Botnet attack trough SSL/TLS channel.

Despite the advancement of Botnet technology and fast evolution, research in finding the solution for Botnet detection is still in its infancy because existing studies remain somewhat limited in scope and do not generally include recent research and development (Silva et al. 2012) [13]. Thus, this indicates that there is still room for improvement in Botnet detection, uniquely encrypted Botnet. Many Botnet detection techniques are based on payload analysis, and these techniques, unfortunately, are inefficient for encrypted C&C channels (Shanti & Seenivasan, 2015) [14]. Zhang et al. (2013) [15] prove that Botnet detection techniques that rely on payload analysis could be foiled by encryption. Payload-based analysis requires decryption, and this leads to a privacy issue. Zhao et al. (2013) [16] state that several challenges in Botnet detection remain unaddressed, such as the ability to design detectors which can cope with new forms of Botnet, therefore they proposed the use of machine learning techniques which proven to increase detection accuracy even for dynamic forms of Botnet. Furthermore, Botnet detection approaches for encrypted traffic were not well established, for instances limited signatures, limited features extracted, limited Botnet detected, and insufficient alarm mechanism (Zhao et al., 2013 [16]; Bortolameotti, 2014 [17]; Larinkoski, 2016 [18]).

The purpose of the study is to propose an approach to detect Botnet in the encrypted channel. The solution was devised to secure the gaps in encrypted Botnet detection system especially for the Botnet detection that base on payload analysis. This study will benefit the system administrator as the detection system assist them in monitoring and protecting system security. This research tends to explore the potential of machine learning techniques which expected to produce a detection system with high accuracy, fast detection and autonomous. The autonomous feature provides minimum supervision and self-learning. The findings also will benefit researchers in this area as it opens up to the exploration of the possible machine learning techniques in developing an effective and efficient Botnet detection system.

The implementation of machine learning in Botnet detection is compelling to overcome the limitation in Botnet detection. Cha & Kim (2017) [19] state that despite the limitations of encrypted Botnet C&C detection, machine learning is a promising approach to detect encrypted Botnet C&C communication. In practice, there are several machine learning techniques implemented in Botnet detection system (Bilge et al. 2011 [20]; Chandhankhede 2013 [21]; Guntuku et al. 2013 [22]; Roshna & Ewards 2013 [23]; Hyslip & Pittman 2015 [24]; Ritu & Kaushal 2015 [25]); however there are still gaps to be fulfilled for the research in detecting encrypted Botnet such as inadequacy of detection features used. In order for the system to achieve a high detection rate and fast detection, it requires techniques that offer high precision and fast pattern recognition capabilities. Autonomous in the detection system requires decision support, situation awareness, and knowledge management. Therefore, this research tends to look at any potential machine learning techniques to fulfill the detection system requirements. In general, machine learning has been proven by previous researches as being able to solve issues like accuracy (Salvador et al. 2009 [26]; Al-Hammadi 2010 [27]; Bilge et al. 2011 [20]; Guntuku et al. 2013 [22]; Ritu & Kaushal 2015 [25]) and real-time (Salvador et al. 2009 [26]; Guntuku et al. 2013 [22]) in Botnet detection.

We organized the remainder of this paper as follows. In section 2, we provide the related work of encrypted Botnet C&C detection. Then in section 3, we propose the Botnet detection framework to detect Botnet over the encrypted channel. This framework will be used to devise the methods based on several phases. Section 4 highlights the challenges for the implementation of the proposed solution. Finally, we drew some concluding remarks in section 5.

II. Related Work

Many researchers relied on a payload-based analysis (deep packet inspection) to detect encrypted Botnet C&C. For example, Zhang et al. (2013) [15] develop high entropy detectors and analyzed packets based on the determined threshold. They stated that the encrypted Botnet produces high entropy, and it can be detected by using the detectors. The challenge of this approach is how to differentiate entropy produces by encrypted Botnet with other traffic that produces high entropy, for instance, media, executable, and compressed files. Tyagi et al. (2015) [28] also implement deep packet inspection (DPI) in their approach and proposed N-gram based HTTP bot traffic detection. The proposed technique detects encrypted and regular Botnet. This technique was based on the fact that the C&C responds with similar communication patterns, with only slight modifications to an HTTP GET request made by a bot. The communications patterns did not varied unless the bot is updated. Therefore this technique works appropriately only if the bot is not updated.

Other work on deep packet inspection was by Sherry et al. (2015) [29], which propose BlindBox to perform deep packet inspection directly on the encrypted traffic. They demonstrate that BlindBox enables applications such as IDS, exfiltration detection, and parental filtering and supports real rule sets from both open-source and industrial DPI systems. They also implement BlindBox and show that it is practical for settings with long-lived HTTPS connections. However, this approach is not specially designed to detect Botnet over the encrypted channel but only stated Botnet as one of their possible usage scenarios. Therefore there was the possibility that this approach might not work well for Botnet detection. Differently, Burghouwt (2015) [6] uses a Causal analysis of traffic flows to detect covert Botnet, for example, Botnet that hides in the encrypted channel. This approach detected covert Botnet by identifying the direct Causal relationship between network flows and prior events. However, this technique needs user events in addition to network traffic; therefore, it causes deployment complexity. Another researcher that used this method is Zhang (2017) [5].

Instead of deep packet inspection, some researchers use decryption techniques to detect encrypted Botnet C&C, for example, Rossow & Dietrich (2013) [12]. They propose PROVEX, a system that automatically derives probabilistic vectorized signatures. PROVEX learns characteristic values for fields in the C&C protocol by evaluating byte probabilities in C&C input traces used for training. This way, they identify the syntax of C&C messages without the need to specify C&C protocol semantics manually, but purely based on network traffic. Even though PROVEX can detect all studied malware families, the fact that it used payload-based analysis that depends on the studied signature limits the detection to the known bots only. Furthermore, by implementing a brute-force-like decryption technique, it leads to the privacy issue.

Some researchers claim that their general Botnet detection approaches could even detect encrypted Botnet C&C based on the assumption that their approaches did not analyze the payload content, and it was Botnet structure independent. For example, Shin et al. (2012) [30], Khan et al. (2015) [31], and Shanti & Seenivasan (2015) [14] commonly use traffic flow analysis in their works. Consequently, they did not have to inspect the payload. However, only Shin et al. provide the detection result of encrypted Botnet C&C even though it was not that prominent. On the other hand, Shanti & Seenivasan provide Botnet detection results in general.

Many researchers implement a machine learning algorithm (MLA) and data mining techniques in their proposed approaches, which can detect Botnet over the encrypted channel. For example, Warmer (2011) [32], Dietrich et al. (2013) [33], Tegeler et al. (2012) [34], Bortolameotti (2014) [17], Wang (2014) [35], Buriya et al. (2015) [36], Cha & Kim (2016) [19] and Jianguo et al. (2016) [37] are some of the researchers that leverage machine learning. Even so, some of them also use deep packet inspection, for instance, Tegeler et al., Wang, and Cha & Kim. Tegeler et al. propose BOTFINDER that uses MLA to identify the key features of Botnet based on the observing traffic that bots produced in the controlled environment. Cha & Kim propose a machine learning approach with several randomness tests to achieve high accuracy detection of encrypted traffic while requiring low overhead incurred by the detection procedure. They test their approach using four MLAs for classification and recommence CART which produced 99.9% accuracy and 2.9 times more efficient than second-best MLA (Naïve Bayes). Wang proposes a novel meta-level classification algorithm based on content features and flow features of traffic. Then he use Naive Bayes classification algorithms to detect encrypted Botnet traffic.

Saad et al. (2011) [38] study the ability of five different commonly used MLAs to meet online Botnet detection requirements, namely adaptability, novelty detection, and early detection. All five MLAs provide high true positive value; however Support Vector Machine got the highest true positive value, which is 97.8%. Warmer (2011) [32] compare different techniques and based on the result proposes three new techniques for detecting HTTP and HTTPS-based C&C channel. It shows Naïve Bayes got the highest true positive which is 97.3%. Ritu & Kaushal (2015) [25] compare different supervised MLAs for determining peer to peer Botnet detection accuracy. Decision Tree and Support Vector Machine achieved 100% accuracy.

Shanti & Seenivasan (2015) [14] propose a detection methodology to classify bot hosts from the normal host by analyzing traffic flow characteristics based on time intervals instead of payload inspection. They use the Decision Tree and Naïve Bayes classification. Classification with a decision tree gave a better true positive of 86.69%. Kirubavathi & Anitha (2016) [39] propose an approach to detect Botnet irrespective of their structures. They try several MLAs to their approach, and Naïve Bayes has the highest detection rate of 99.14%.

Zhao et al. (2013) [16] and Bortolameotti (2014) [17] use Decision Tree to their approaches, and both provide a very high detection rate which is 98.5 % and 99.96% with a very low false positive rate of 0.01 % and 0%. Dietrict et al. (2013) [33] develop CoCoSpot use Average-Linking Hierarchical Clustering. 50% of Botnet families were detected by the rate of 95.6%. Buriya et al. (2015) [36] use Naïve Bayes and achieved 98.84% accuracy. Apparently most of the MLAs discussed in this paper have a very high detection rate.

Even though some techniques provide a high detection rate, comparatively they also got a high false positive rate. For example, Richer (2017) [40] proposes an approach using Support Vector Machine and got a 100% detection rate; however the false positive rate is more than 15%. The work by Shanti & Seenivasan (2015) [14] also provides very high false positive which is more than 21%. Above all, Warmer (2011) had the highest false positive value of 44.3% by using Naïve Bayes.

Al-Hammadi (2010) [27] presents a host-based behavioral approach for detecting Botnet based on correlating different activities generated by bots by monitoring function calls within a specified time window. Al-Hammadi uses Dendric Cell Algorithm inspired by the Immune System. The evaluation shows that correlating different activities generated by IRC/P2P bots within a specified period achieves high detection accuracy (100%). In addition, using an intelligent correlation algorithm not only states if an anomaly is present, but it also exposed the source of the anomaly.

One of the most prominent MLA for Botnet detection is Neural Network and its distributions, Self-Organizing Map (SOM). SOM is an unsupervised Neural Network and has been widely used in intrusion detection. Unfortunately, there are limited works discussing SOM for Botnet detection, and instead, more in intrusion detection. However, SOM is a promising approach especially for developing an autonomous Botnet detection system. Langin et al. [i] (2009) [41] use SOM to cluster and classify peer to peer Botnet traffic and other malignant network activity by analyzing firewall log entries. Langin et al. [ii] (2009) [42] use Hexagonal SOM for clustering and then for classification of new firewall log data to look for new bots in the network.

Guntuku et al. (2013) [22] propose and implement a hybrid framework for detecting peer to peer Botnet in live network traffic by integrating Neural Networks with Bayesian Regularization for the detection of newer and unseen Botnet in live traffic of a network. It was conclusively shown through the statistical tests that the trained Bayesian Regularization - Neural Network model can generalize very well and can predict the activity of unknown bots' malicious activity. Thus Botnet detection activities successfully achieved with an accuracy of 99.2%. Nogueira et al. (2010) [43] extend the framework propose by Salvador and develop the Botnet Security System called BoNeSSy. Nogueira et al. develop a Botnet detection system that is based on the collection of flow statistics using Neural Network. The results obtained show that the system is feasible and efficient since it provides high detection rates with low computational overhead.

Many existing approaches to the process of detecting intrusions utilized some forms of rule-based analysis. Expert System is the most common form of rule-based intrusion detection approaches. Most existing behavior-based approaches are not able to detect and predict the Botnet as they change their structure and pattern. Roshna & Ewards (2013) [23] present the AdaptiveNeuro-Fuzzy Inference System (ANFIS), a technique that trains the system for future prediction. However, the limitation of this work is the restriction of fuzzy rules and fuzzy sets for the comparison purpose. Therefore, the proposed work should be able to overcome the limitations by increasing the number of rules generated using the Botnet features and information gain.

Fuzzy pattern recognition proposed by Wang et al. (2011) [44] intends to identify bot-relevant domain names and IP addresses by inspecting network traces. The algorithm involves traffic reduction, feature selection, and pattern recognition. Fuzziness in pattern recognition helps to detect bots that are hidden or camouflage. Performance evaluation results based on real traces show that the proposed system can reduce more than 70% input raw packet traces and achieve a high detection rate (about 95%) and a low false positive rate (0–3.08%). Furthermore, the proposed FPRF algorithm is resource-efficient and can identify inactive Botnet to indicate potentially vulnerable hosts. BotDigger proposed by Al-Duwairi &Al-Ebbini (2010) [45] utilizes fuzzy logic in order to define logical rules that are mainly based on some statistical facts and essential features that identify Botnet activities. The key advantage of the architecture designed in this research is that it allows the integration of a wide range of traffic specifications.

The above machine learning approaches mostly use detection rate, accuracy, or false positive value as the metrics to measure the detection performance. However, other vital metrics are real-time and autonomous. Even though detection able to detect accurately, it is useless without fast detection or real-time detection. Researchers focus on developing a real-time Botnet detection system, for example, Salvador et al. (2009) [26], Wang et al. (2011) [44] and Guntuku et al. (2013) [22].

Autonomous mainly focus on self-learning and self-managing properties. Chandhankhede (2013) [21] proposes the new autonomous model for Botnet detection using the K-means algorithm, one of the most straightforward unsupervised learning algorithms that solve the well-known clustering problem. According to Khattak et al. (2014) [46], the degree of automation can be classified as manual, semi-automated and automated. Semi-automated Botnet detection requires very little human intervention, and most of the detection is performed on automated fashion. However, fully automated Botnet detection should require no human intervention after initial development. Khattak et al. also agree that ideally, any detection method should be as generic and automated as possible.

III. Proposed Botnet Detection Framework

To achieve high accuracy, fast detection, and an autonomous Botnet detection system as stated in section I, we propose a conceptual framework as a guideline to devise a methodology to detect Botnet over the encrypted channel. Fig. 2 shows the Botnet Analysis and Detection System (BADS) framework which consists of three main components namely Network Analysis System (NAS), IDS and Alarm System (AS). Through BADS, we generally divide the process into six phases as depicted in Fig. 3 also shows the expected results of each phase.

Phase 1: Data Collection, Customization and Conversion

This study requires the dataset of encrypted Botnet traffic; however, in the secure network, it is a challenge to get one. Therefore, we convert public Botnet datasets into customizable encrypted Botnet dataset by using BotTalker developed by Zhang & Papadopoulos (2013) [47]. The public datasets use are ISOT, Malware Capture Facility Project (MCFP), and Network Information Management and Security (NIMS). Some datasets are in pcap format; therefore, we need to convert the dataset into CSV format using Wireshark, then to arff using Weka.

Phase 2: Feature Extraction and Selection

In encrypted Botnet, detection by inspecting the payload is a tedious process, especially if it involves extensive traffic data. Furthermore, this method required the decryption of data, and it involved a privacy issue. Because of that, for this research, the approach without inspecting the payload is necessary. Encrypted Botnet itself produces features that can be used for detection, and most importantly, it should not require any decryption. Botnet features are extracted through a feature extraction process, and then if necessary, followed by feature selection. Feature selection reduces the number of features, and these selected features are the most relevant features for the detection. Feature selection is crucial because it is a reliant factor to detection accuracy, as proven by Buriya et al. (2015) [36] and Kirubavathi & Anitha (2016) [39].

Tranalyzer is used to extract the Botnet features. From the literature study, Tranalyzer was proven to capture more features compared to other features extractors. Furthermore, features extracted using Tranalyzer provide better accuracy (Jianguo et al., 2016) [37]. For feature selection, we use Information Gain Attribute Evaluation in Weka and employ Ranker Algorithm to select the features that will give the most relevant features based on the ranking provided.

Phase 3: Botnet Prediction and Classification

We use Weka classify module to perform Botnet prediction and classification. The classification generates decision rules, and these rules are used for Botnet detection.

Phase 4: Botnet Detection

IDS component consists of a Snort-based Botnet detection mechanism, as shown in Fig. 4. Sensors sniff the packets from the network. Packet decoder takes packets from different types of network interfaces and prepares the packets to be preprocessed. The preprocessor arranges or modifies data packets before the detection engine does some operation. It also normalizes protocol headers, detects anomalies, packet re-assembly, and TCP stream reassembly. The detection engine is an essential part of IDS. The detection engine detects Botnet intrusion activity that exists in a packet. There are two detectors use; misuse detector and anomaly detector. The detection engine employs fuzzy inference rules for this purpose. The rules are read into internal data structures or chains where they are matched against all packets. If a packet matches any rule, appropriate action is taken; otherwise, the packet is dropped. The idea of fuzzy rules and fuzzy inference implementation in the detection engine is to determine the severity level of a Botnet attack.

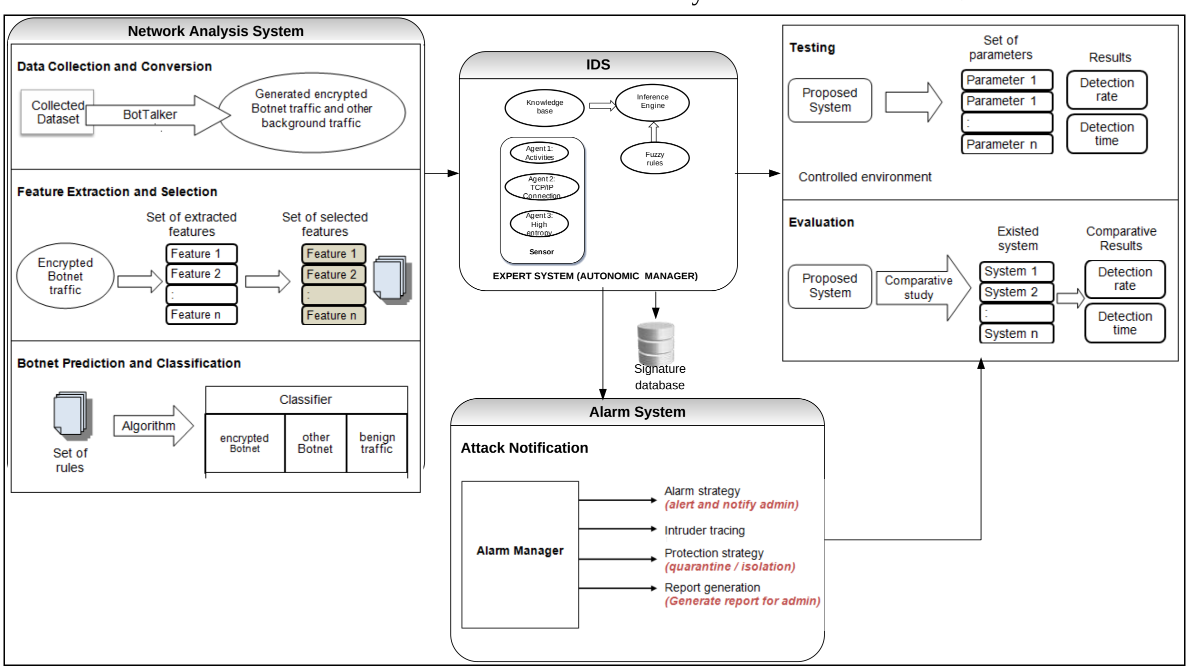


Fig. 2. Botnet Analysis and Detection System (BADS) Framework.

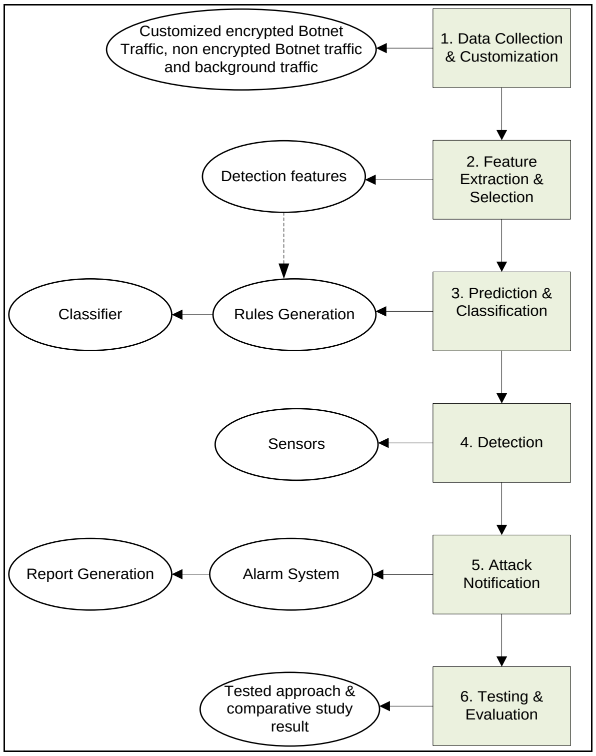


Fig. 3. Phases in Detecting Encrypted Botnet.

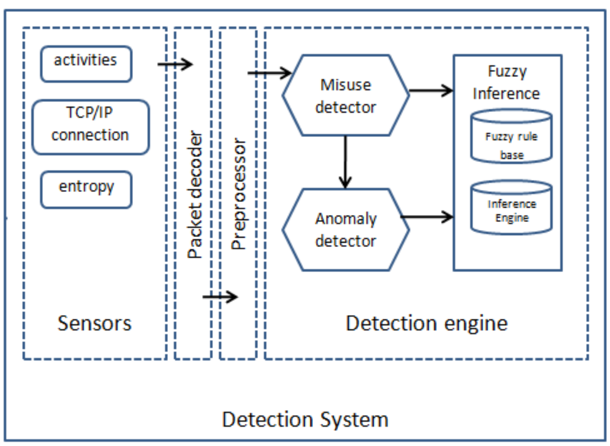


Fig. 4. Fuzzy Inference Snort-Based Detection System.

Phase 5: Attack notification

Intrusion notification assists the network manager to manage the system. There are four sub-components to fulfill that purpose namely intruder tracing, alarm strategy, protection strategy, and report generation. Implementing these components into the system should be able to alert the network manager and notify them of the severity of attacks, suggesting protection strategy, and generate a report. The autonomous mechanism enables the Botnet detection system to works effectively with minimum human intervention.

Phase 6: Testing and Evaluation

The open stack test-bed is set up by using multiple virtual machines. Several virtual machines are used to carry out intrusion attempts and one virtual machine to run the proposed Botnet detection system. This phase happens to test the effectiveness of the proposed approach by using several parameters, for instance, accuracy and false positive. Then, for the evaluation, the comparison study is performed between the proposed approach and other existing Botnet detection system to measure the efficiency of the approach.

The BADS assists the network manager in monitoring the security of the system. The idea of BADS is to minimize human intervention in performing network monitoring and suppose to take appropriate action based on the severity of the Botnet attack. Therefore we endeavor to propose an autonomous Botnet detection system.

IV. Challenges of Implementation

There are several challenges to implement BADS. Firstly, it is data preprocessing parts, which are dataset collection, dataset customization, and conversion. Actually, for this work, we also want to use the Botnet dataset from IMPACT Cyber Trust. However, the dataset is only available to US-based researchers and those in approved locations. Unfortunately, our location is not in that approved locations. Another challenge is for data customization as the reference for BotTalker is quite limited. Then, we have to do data conversion for all the datasets except for the NIMS dataset. Overall, the data preprocessing part is a tedious part and requires many works.

Since we are using various tools in our work, we are expecting conflict because each tool produces different types of outputs. For example, the rules that are retrieved from the classification in Weka to the rules structure in Snort. Furthermore, we also want to employ fuzzy rules into Snort because we want the detection system to be able to determine the severity of Botnet attacks. We try to achieve this because we want the detection system to provide appropriate solutions based on the level of a Botnet attack. This feature will help the network manager to monitor the network and provides automation. We believe fuzzy inference rules can provide the required solution. Currently, we are still looking for solutions to this issue.

V. Conclusion and Future Work

Botnet evolves, and new Botnet strains have developed advanced evasion techniques. It includes the capabilities to manipulate encrypted channels like SSL/TLS for their command and control communication, use social media to spread malware, spamming, and gain credential info (social bot). These avoidance techniques enable Botnet to cover its operation, evade detection, and stay on the system as long as possible. However, existing detection techniques were not well established and had limitations in detecting Botnet especially the Botnet over the encrypted channel. Having an effective and efficient Botnet detection system is essential. This research endeavors to find a solution to enhance the Botnet detection system over the encrypted channel by using machine learning. Machine learning is a promising approach to detect Botnet, especially over an encrypted channel. Therefore, we proposed the BADS framework and devised a methodology based on the framework.

This framework consists of three main components, which are Network Analysis System (NAS), IDS, and Alarm System (AS). Besides the main components, testing and evaluation processes also included in the framework. We devise the methodology from the framework and divide them into six phases. Overall the contribution of this paper is three-fold:

1) The framework of Botnet Analysis and Detection System (BADS).

2) The methodology of devising the techniques for Botnet detection.

3) The design of fuzzy inference Snort-based Botnet detection system.

ACKNOWLEDGMENT

This research is supported by the Universiti Sains Malaysia Research University (RUI) Grant (account number 1001/ PKOMP/ 8014017).

REFERENCES

[1] Soloman, A. & Evron, G., The World of Botnet, Virus Bulletin, September 2006, www.beyondsecurity.com/whitepapers/SolomonEvronSept06.pdf.

[2] Security Press Release, Botnets Threaten National Infrastructure and Security, 24 October 2006, http://www.itsecurity.com/press-releases/press-release-mcafee-botnets-102506/.

[3] Thomas, V. & Jyoti, N., Virus Bulletin, Defeating IRC Bots on the Internal Network, 1 February 2007, https://www.virusbulletin.com/virusbulletin/2007/02/defeating-irc-bots-internal-network.

[4] MyCERT Incident Statistics, https://www.mycert.org.my/statistics/2018.php.

[5] Zhang, H., Detecting Advanced Botnets in Enterprise Networks, Ph.D. Thesis, 2017, Department of Computer Science, Colorado State University, USA.

[6] Burghouwt, P., Detection of Botnet Command and Control Traffic in Enterprise Networks, Ph.D. Thesis, 2015, The Hague University of Applied Science, Netherlands.

[7] Nicholson, P., What Lies Beneath: Advanced Attacks that Hide in SSL Traffic, September 30, 2015, https://www.a10networks.com/blog/what-lies-beneath-advanced-attacks-hide-ssl-traffic.

[8] Paganini, P., SSL Blacklist A New Weapon to Fight Malware and Botnet, http://securityaffairs.co/wordpress/26672/cyber-crime/ssl-blacklist-new-weapon-fight-malware-botnet.html, 2014 access on 3 August 2017.

[9] Gebhart, G., We Are Halfway to Encrypting The Entire Web, Electronic Frontier Foundation (EFF) Report, 2017, https://www.eff.org/deeplinks/2017/02/were-halfway-encrypting-entire-web.

[10] Finley, K., Half of the Web is Now Encrypted. That Makes Everyone Safer, 2017, https://www.wired.com/2017/01/half-web-now-encrypted-makes-everyone-safer/.

[11] Gooley, D., The Rise in SSL-based Treats, 2017, https://www.zscaler.com/blogs/research/rise-ssl-based-threats.

[12] Rossow, C. & Dietrich, C.J., (2013). ProVeX: Detecting Botnets with Encrypted Command and Control Channels, International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA), 18-19 July 2013, Berlin, Germany, pp 21-40.

[13] Silva, S. S. C., Silva, R. M. P., Pinto, R. C. G., & Salles, R. M. (2013). Botnets: A survey. Computer Networks, 57(2), 378–403. https://doi.org/10.1016/j.comnet.2012.07.021.

[14] Shanti, K. & Seenivasan, D., (2015). Detection of Botnet by Analyzing Network Traffic Flow Characteristics using Open Source Tools, 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO), 9-10 January 2015, Andhra Pradesh, India, DOI: 10.1109/ISCO.2015.7282353.

[15] Zhang, H., Papadopoulos, C., & Massey, D., (2013). Detecting Encrypted Botnet Traffic, 2013 IEEE INFOCOM, 14-19 April 2013, Turin, Italy, pp 3453-3458.

[16] Zhao, D., Traore, I., Sayed, B., Lu, W., Saad, S., Ghorbani, A., & Garant, D. (2013). Botnet Detection based on Traffic Behavior Analysis and Flow Intervals. Computers and Security, 39(PARTA), 2–16. https://doi.org/10.1016/j.cose.2013.04.007.

[17] Bortolameotti, R. (2014). C&C Botnet Detection over SSL, Master Thesis, University of Twente, Netherlands.

[18] Larinkoski, L. (2016). Detecting Encrypted C&C using Network Fingerprints

[19] Cha, S., & Kim, H. (2017). Detecting Encrypted Traffic: A Machine Learning Approach (pp. 54–65). Springer, Cham. https://doi.org/10.1007/978-3-319-56549-1_5.

[20] Bilge, L. (2011). EXPOSURE : a Passive DNS Analysis Service to Detect and Report Malicious Domains, V(4).

[21] Chandankhede, P. (2013). Autonomous Botnet Detection, *3*(13), 71–76

[22] Guntuku, S., Narang, P., & Hota, C. (2013). Real-time Peer-to-Peer Botnet Detection Framework based on Bayesian Regularized Neural Network. ArXiv Preprint ArXiv:1307.7464. Retrieved from http://arxiv.org/abs/1307.7464.

[23] Roshna, R. S., & Ewards, V. (2013). Botnet Detection Using Adaptive Neuro Fuzzy Inference System, *3*(2), 1440–1445.

[24] Hyslip, T., & Pittman, J. (2015). A Survey of Botnet Detection Techniques by Command and Control Infrastructure. Journal of Digital Forensics, Security and Law, 10(1), 7–26. https://doi.org/10.1145/1090191.1080118.

[25] Ritu & Kaushal R. (2015), Machine Learning Approach for Botnets Detection, 3rd Security and Privacy Symposium, 13-14 February 2015, IIIT – Delhi..

[26] Salvador, P., Nogueira, a., Franca, U., & Valadas, R. (2009). Framework for Zombie Detection Using Neural Networks. 2009 Fourth International Conference on Internet Monitoring and Protection, 14–20. https://doi.org/10.1109/ICIMP.2009.10.

[27] Al-Hammadi, Y. A. A. (2010). Behavioural correlation for malicious bot detection. Doctor, (April). Retrieved from http://etheses.nottingham.ac.uk/1359/.

[28] Tyagi, R., Paul, T., Manoj, B. S., & Thanudas, B. (2015). A novel HTTP botnet traffic detection method. 12th IEEE International Conference Electronics, Energy, Environment, Communication, Computer, Control: (E3-C3), INDICON 2015, 1–6. https://doi.org/10.1109/INDICON.2015.7443675.

[29] Sherry, J., Lan, C., Popa, R. A., & Ratnasamy, S. (2015). BlindBox: Deep Packet Inspection over Encrypted Traffic. Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication - SIGCOMM '15, 213–226. https://doi.org/10.1145/2785956.2787502.

[30] Shin, S., Xu, Z., & Gu, G. (2012). EFFORT: Efficient and effective bot malware detection. Proceedings - IEEE INFOCOM, 2846–2850. https://doi.org/10.1109/INFCOM.2012.6195713.

[31] Khan, A., Ahlawat, C., & Bijalwan, A. (2015). A unified botnet detection framework 1, International Journal of Advances in Electronics and Computer Science (March), Vol. 2, Issue 5, pp 81-87.

[32] Warmer, M. (2011). Detection of web-based command & control channels, (November). Master Thesis, University of Twente, Netherlands. Retrieved from http://essay.utwente.nl/61232/.

[33] Dietrich, C. J., Rossow, C., Pohlmann, N. (2013), CoCoSpot: Clustering and Recognizing Botnet Command and Control Channels using Traffic Analysis, Journal of Computer Networks, Vol. 57, Issue 2, pp. 475-486.

[34] Tegeler, F., Fu, X., Vigna, G., & Kruegel, C. (2012). BotFinder: Finding Bots in Network Traffic Without Deep Packet Inspection. Proceedings of the 8th International Conference on Emerging Networking Experiments and Technologies - CoNEXT '12, 349. https://doi.org/10.1145/2413176.2413217.

[35] Wang, Y. (2014). Encrypted botnet detection scheme. Proceedings - 2014 9th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC 2014, 559–565. https://doi.org/10.1109/3PGCIC.2014.110.

[36] Buriya, S., Patel, A. K., Yadav, S. S., Buriya, S., Patel, A. K., & Yadav, S. S. (2015). Botnet Behavior Analysis Using Naïve Bayes Classification Algorithm Without Deep Packet, *IX*(Viii), 45–54.

[37] Jianguo, J., Qi, B., Zhixin, S., Wang, Y., & Lv, B. (2016). Botnet Detection Method Analysis on the Effect of Feature Extraction, 1884–1890. https://doi.org/10.1109/TrustCom.2016.286.

[38] Saad, S., Traore, I., Ghorbani, A., Sayed, B., Zhao, D., Lu, W., … Hakimian, P. (2011). Detecting P2P botnets through network behavior analysis and machine learning.2011 Ninth Annual International Conference on Privacy Security and Trust, 174–180. https://doi.org/10.1109/PST.2011.5971980.

[39] Kirubavathi, G., & Anitha, R. (2016). Botnet detection via mining of traffic flow characteristics. Computers and Electrical Engineering, 50, 91–101. https://doi.org/10.1016/j.compeleceng.2016.01.012.

[40] Richer, T. J. (2017). Entropy-based detection of botnet command and control. Proceedings of the Australasian Computer Science Week Multiconference on - ACSW '17, 1–4. https://doi.org/10.1145/3014812.3014889.

[41] Langin, C., Zhou, H., & Rahimi, S. (2009). A Self-Organizing Map and its Modeling for Discovering Malignant Network Traffic, (Mar). https://doi.org/10.1109/CI-CYBS.2009.4925099.

[42] Langin, C., Che, D., Wainer, M., & Rahimi, S. (2009). Visualization of Network Security Traffic using Hexagonal Self-Organizing Maps.

[43] Nogueira, A., Salvador, P., & Blessa, F. (2010). A Botnet Detection System Based on Neural Networks. Digital Telecommunications (ICDT), 2010 Fifth International Conference On, 57–62. https://doi.org/10.1109/ICDT.2010.19.

[44] Wang, K., Huang, C. Y., Lin, S. J., & Lin, Y. D. (2011). A fuzzy pattern-based filtering algorithm for botnet detection. Computer Networks, 55(15), 3275–3286. https://doi.org/10.1016/j.comnet.2011.05.026.

[45] Al-Duwairi, B., & Al-Ebbini, L. (2010). BotDigger: A fuzzy inference system for botnet detection. 5th International Conference on Internet Monitoring and Protection, ICIMP 2010, 16–21. https://doi.org/10.1109/ICIMP.2010.11

[46] Khattak, S., Ramay, N. R., Khan, K. R., Syed, A. A., & Khayam, S. A. (2014). A Taxonomy of Botnet Behavior ,*16*(2), 898–924.

[47] Zhang, H., Papadopoulos, C., & Massey, D. (2013). Detecting encrypted botnet traffic. Proceedings - IEEE INFOCOM, 3453–3458. https://doi.org/10.1109/INFCOM.2013.6567180.

The Multi-Class Classification for the First Six Surats of the Holy Quran

Nouh Sabri Elmitwally[1], Ahmed Alsayat[2]
Department of Computer Science, College of Computer and Information Sciences, Jouf University, Aljouf, Saudi Arabia[1, 2]
Department of Computer Science, Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt[1]

***Abstract*—The Holy Quran is one of the holy books revealed to the prophet Muhammad in the form of separate verses. These verses were written on tree leaves, stones, and bones during his life; as such, they were not arranged or grouped into one book until later. There is no intelligent system that is able to distinguish the verses of Quran chapters automatically. Accordingly, in this study we propose a model that can recognize and categorize Quran verses automatically and conclusion the essential features through Quran chapters classification for the first six Surat of the Holy Quran chapters, based on machine learning techniques. The classification of the Quran verses into chapters using machine learning classifiers is considered an intelligent task. Classification algorithms like Naïve Bayes, SVM, KNN, and decision tree J48 help to classify texts into categories or classes. The target of this research is using machine learning algorithms for the text classification of the Holy Quran verses. As the Quran texts consists of 114 chapters, we are only working with the first six chapters. In this paper, we build a multi-class classification model for the chapter names of the Quranic verses using Support Vector Classifier (SVC) and GaussianNB. The results show the best overall accuracy is 80% for the SVC and 60% for the Gaussian Naïve Bayes.**

Keywords—Text classification; machine learning; natural language processing; text pre-processing; feature selection; data mining; Holy Quran

I. Introduction

Text classification of the Holy Quran is a research topic researchers should pay attention to in the context of machine learning algorithms.

The Holy Quran is a book that was sent down from the heavens into the heart of the prophet Muhammad to be delivered to all human beings, not only Muslims. The sacred words were revealed by Allah and written into a meaningful textual format that could be analysed and classified using machine learning classification algorithms.

It is considered a comprehensive book covering every component of life and accessible to all people. It addresses the heart and mind as one.

The texts of the Holy Quran are fertile ground for natural-language processing and text classification. Their uniqueness and meanings distinguish the features. The Holy Quran is the first source of legislation in Islam. It is necessary to apply data-mining techniques to classify the verses into chapters (surats) intelligently based on machine learning techniques.

Furthermore, annotation of the verses of the Holy Quran's surats depends not only on the text itself but also on the ordering of the surats. Therefore, this study builds a model to classify and differentiate Quranic verses, according to their surats.

We have previously studied the architecture of the Arabic Language Sentiment Analysis (ALSA) [1]. We extended the concept of text classification to apply it to the Holy Quran's verses. The total number of verses in the Holy Quran is about 6000. Multi-class classification means that we need an automating model that enables classification of the texts accordingly. For this reason, this paper looks at the first six chapters from the Holy Quran; its approximately 1000 verses contain a total 8000 features for the training and testing data.

This paper is constructed as follows: the next section presents related work on multi-class text classification of the Holy Quran. Experimental method and analysis are covered in Section 3. Finally, the fourth section includes the results followed by the conclusions and anticipations of future work.

II. Related Work

The study detailed in [2] proposed an automation model that could classify Al-hadeeth features into Sahih, Hasan, Da'if, and Maudu, using machine learning techniques (LinearSVC, SGDClassifier, and LogisticRegression).

The author of [3] built a machine-learning model using an algorithm (KNN, SVM, and Naïve Bayes) classification model to annotate labels for the Quranic verses. The accuracy of the text-classification algorithms reached over 70% for the multi-labels of the Quranic verses.

The authors of [4] proposed a multi-label classification approach to the topics of Quranic verses using a k-Nearest Neighbor (KNN) algorithm with a weighted TF-IDF and TF-IDF.

Another research paper looked at the impact evaluation for four classification algorithms (SVM, KNN, Naïve Bayes and Decision Tree) to classify the topic of the Quranic Ayāts/verses [5]. The same concept as studied in [6] used the MultinomialNB classifier.

The authors of [7] used the Propbank Corpus to improve the performance of semantic argument classification on Quran data using the SVM Linear.

The authors of [8] applied the GBFS approach to label Quranic verses based on two major references, the commentary on the verses and the English translation. In addition, they proposed the IG-CFS technique to label Quranic verses of surats al-Baqara and al-Anaam [9].

III. Experiment and Analysis

The proposed model consists of four important phases as shown in the following framework architecture: 1) data collection, 2) text feature engineering, 3) The Term Frequency – Inverse Document Frequency (TF-IDF) feature representation, and 4) The GaussianNB and SVC classifiers. The framework architecture of the multi-class Quran framework classification is shown in Fig. 1.

A. Data Pre-processing and Cleaning

Before machine-learning modelling, we applied text pre-processing and cleaning techniques to extract features according to the following steps: remove the Arabic Tashkeel symbols (e.g., ◌ٍ◌ِ◌ُ◌َ◌ً); and remove consecutive Tatweel ('ـ') within Arabic characters.

B. Corpus

The corpus size was 954 verses collected from the first six surats of the Holy Quran. Table I shows generated descriptive statistics summarizing the central tendency, dispersion and the shape of the corpus' distribution.

Table II outlines the extracted sample from the Holy Quran corpus for the six classified categories ["Fatiha", "Albaqrah", "AlEimran", "Alnisaa", "Almayida", "Alaneam"] in the first column. The number of verses is shown in the second column. The selected verse and its translation appear in columns three and four.

C. Exploratory Data Analysis

The goal of the Exploratory Data Analysis (EDA) is to extrapolate on the breadth of information reflected by the corpus data. Fig. 2 shows the number of verses per corpus class.

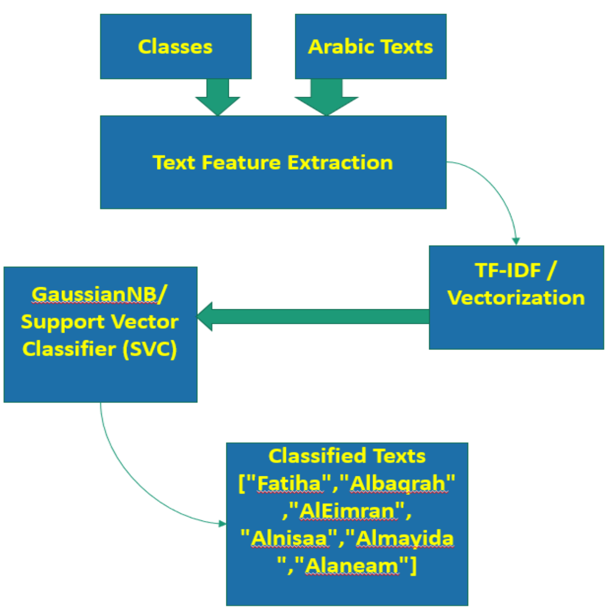


Fig. 1. The Quran Framework Classification.

TABLE I. The Descriptive Summary of the Holy Quran Corpus

count	954.000000
mean	3.640461
std	1.471548
min	1.000000
25%	2.000000
50%	3.000000
75%	5.000000
max	6.000000

TABLE II. Examples of Quran Verses

Chapter/Surat	**Number of verses /Aya**	**Arabic Text**	**English Text**
Fātiha/ the Opening Chapter/الفاتحة	7	بسم الله الرحمن الرحيم	In the name of God, Most Gracious, Most Merciful.
Baqara/ Heifer/البقرة	286	الم ذلك الكتاب لا ريب فيه هدى للمتقين	A.L.M This is the Book; in it is guidance sure, without doubt, to those who fear God
AlEimr Āl-i-'Imrānan/ The Family of 'Imrān/آل عمران	200	الم الله لا اله هو الحى القيوم	Allah! there is no god but He the Living the Self-Subsisting Eternal
Nisāa/The Women/النساء	176	يا أيها الناس اتقوا ربكم الذى خلقكم من نفس واحدة وخلق منها زوجها وبث منهما رجالا كثيرا ونساء واتقوا الله الذى تساءلون به والأرحام ان الله كان عليكم رقيبا	O mankind! reverence your Guardian-Lord Who created you from a single person created of like nature his mate and from them twain scattered (like seeds) countless men and women; reverence God through Whom ye demand your mutual (rights) and (reverence) the wombs (that bore you): for God ever watches over you.
Māida/ The Table Spread/المائدة	120	يا أيها الذين ءامنوا أوفوا بالعقود أحلت لكم بهيمة الأنعام إلا ما يتلى عليكم غير محلى الصيد وأنتم حرم إن الله يحكم ما يريد	O ye who believe! fulfil (all) obligations. Lawful unto you (for food) are all four-footed animals with the exceptions named: but animals of the chase are forbidden while ye are in the Sacred Precincts or in pilgrim garb: for God doth command according to His Will and Plan.
An'ām/Catle/ الأنعام	165	الحمد لله الذى خلق السماوات والأرض وجعل الظلمات والنور ثم الذين كفروا بربهم يعدلون	

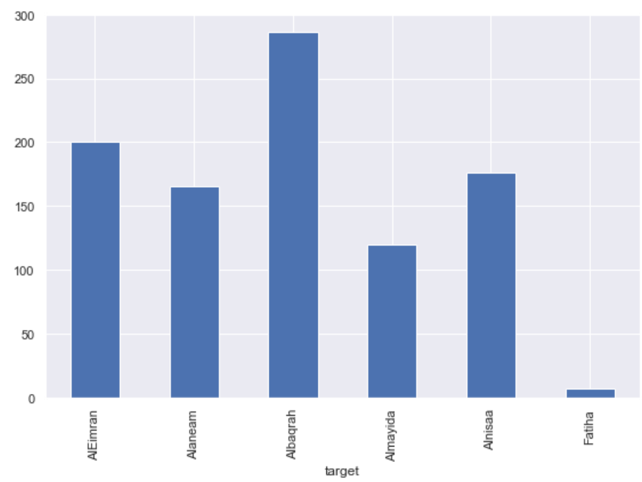


Fig. 2. Number of Verses Per Class.

D. Feature Engineering and Selection

Feature-text selection and engineering are considered the process of choosing the essential features required to represent the model for machine-learning classifiers. The following figures (Fig. 3-8) show word clouds for each Surat in the Holy Quran corpus.

Fig. 3. Word Cloud-Surat Al Fatiha.

Fig. 4. Word Cloud-Surat Albaqara.

Fig. 5. Word Cloud-Surat AlEimran.

Fig. 6. Word Cloud-Surat AlNisaa.

Fig. 7. Word Cloud-Surat AlMaeda.

Fig. 8. Word Cloud-Surat Alanaam.

IV. Results

We calculated Accuracy, Recall and F1-value according to the following mathematical equations:

$$F - value = \frac{2*Accuracy*Recall}{Accuracy+Recall} * 100\% \quad (1)$$

$$Recall = \frac{The\ number\ of\ corrected\ texts\ of\ a\ specific\ class}{The\ number\ of\ texts\ of\ this\ class\ in\ testing\ data} * 100\% \quad (2)$$

$$Accuracy = \frac{The\ number\ of\ corrected\ texts\ in\ a\ specific\ class}{The\ number\ of\ texts\ in\ the\ class} * 100\% \quad (3)$$

A. Machine-Learning Classifiers

The Support Vector Classifier (SVC) is considered the implementation of the Support Vector Machine (SVM) [5] for solving multi-class classification problems. The GaussianNB performs accurate feature-vector classification for the multi-class text problems [10]. We tested the proposed model against the performance metrics. The results are shown in Table III.

The sample texts of misclassified instance-classes are listed in Table IV. The table shows the missed classified text according to the expected and predicted output for the six classes (“Fatiha”–1; “Albaqrah”–2; “AlEimran”–3; “Alnisaa”– 4; “Almayida”–5; “Alaneam”–6).

B. Evaluation Metrics

The classification algorithms need the performance metrics to measure the model accuracy and losses. Fig. 9 shows that most of the performance metrics we used to evaluate the proposed multi-class Quranic model. The performance metrics are: 1) cohen_kappa; 2) log_loss; 3) zero_one_loss; 4) hamming_loss; and 5) Mathews_corrcoef.

The proposed model is evaluated according to two classifiers, SVC [7] and GaussianNB, as shown in Table V and Table VI and the Fig. 10 and Fig. 11. The performance of the proposed model is measured in terms of accuracy, precision, recall, f-measure, AUC, and ROC curves. The SVC classifier had the highest AUC value of 0.97 while the GaussianNB had the AUC value of 0.82 (see Fig. 12 and Fig. 13).

TABLE. III. THE PERFORMANCE METRICS

Metric	SVC	GaussianNB
cohen_kappa_score	0.408	0.395
log_loss	0.000	16.456
zero_one_loss	0.450	0.476
hemming_loss	0.450	0.476
matthews_corrcoef	0.420	0.396

TABLE. IV. THE MISCLASSIFIED INSTANCE-CLASSES

Text	Expected Output	Predicted Output
وقاتلوهم حتى لا تكون فتنة ويكون الدين لله فإن انتهوا فلا عدوان إلا على الظالمين	2	6
يا أيها الذين آمنوا لا تأكلوا الربا أضعافا مضاعفة واتقوا الله لعلكم تفلحون	3	5
واتل عليهم نبأ ابني آدم بالحق إذ قربا قربانا فتقبل من أحدهما ولم يتقبل من الآخر قال لأقتلنك قال إنما يتقبل الله من المتقين	5	2
وإذا سمعوا ما أنزل إلى الرسول ترى أعينهم تفيض من الدمع مما عرفوا من الحق يقولون ربنا آمنا فاكتبنا مع الشاهدين	5	2
إن أول بيت وضع للناس للذي ببكة مباركا وهدى للعالمين	3	2

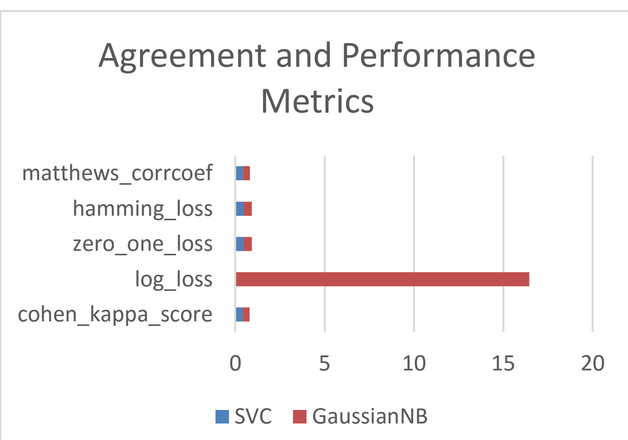


Fig. 9. The Agreement and Performance Metrics.

TABLE. V. RESULTS FOR SVM CLASSIFIER

Class	Precision	Recall	F1- score	Area Under Curve (AUC)
Fatiha	0.000	0.000	0.000	0.80
Albaqrah	0.487	0.475	0.481	0.68
AlEimran	0.545	0.364	0.436	0.85
Alnisaa	0.478	0.754	0.585	0.77
Almayida	0.871	0.771	0.818	0.97
Alaneam	0.444	0.167	0.242	0.76

TABLE. VI. RESULTS FOR GAUSSIANNB CLASSIFIER

Class	Precision	Recall	F1- score	Area Under Curve (AUC)
Fatiha	1.000	0.500	0.667	0.75
Albaqrah	0.424	0.350	0.384	0.61
AlEimran	0.548	0.515	0.531	0.71
Alnisaa	0.550	0.579	0.564	0.69
Almayida	0.686	0.686	0.686	0.82
Alaneam	0.355	0.458	0.400	0.67

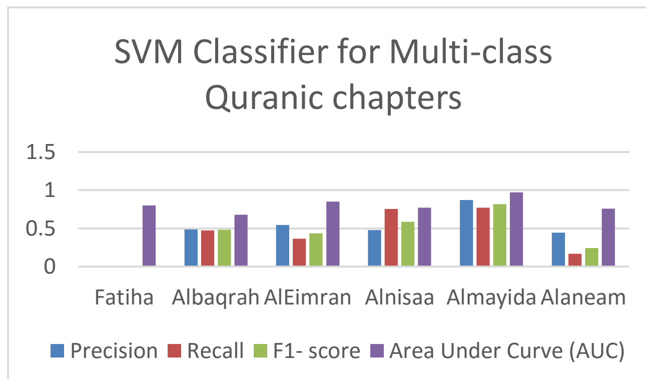


Fig. 10. SVM Classifier for Multi-Class Quranic Chapters.

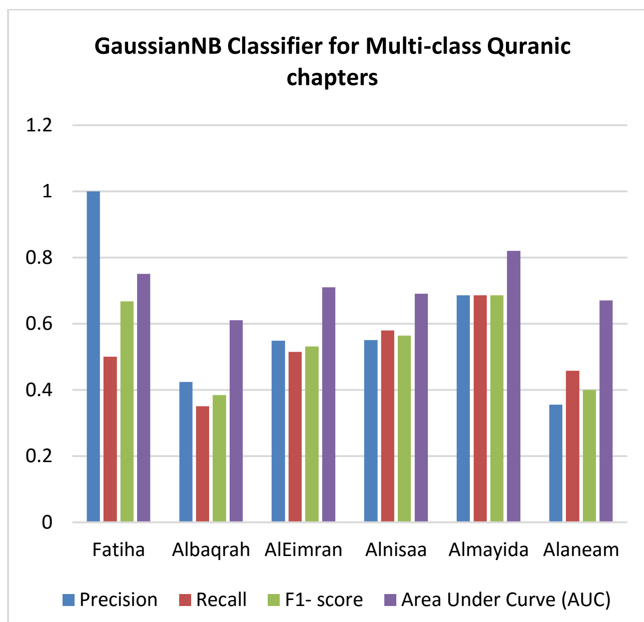


Fig. 11. GaussianNB Classifier for Multi-Class Quranic Chapters.

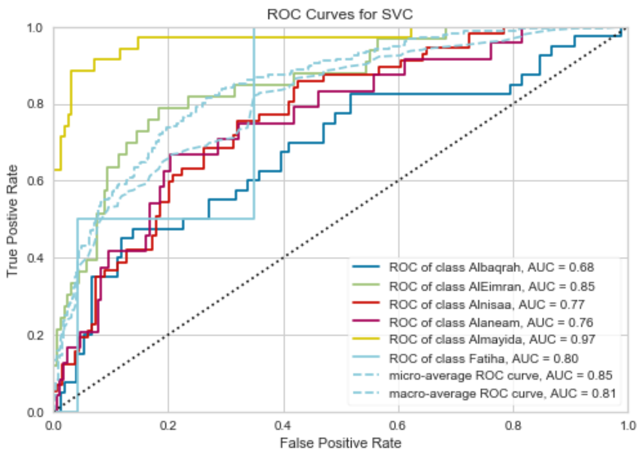


Fig. 12. ROC Curves – SVC.

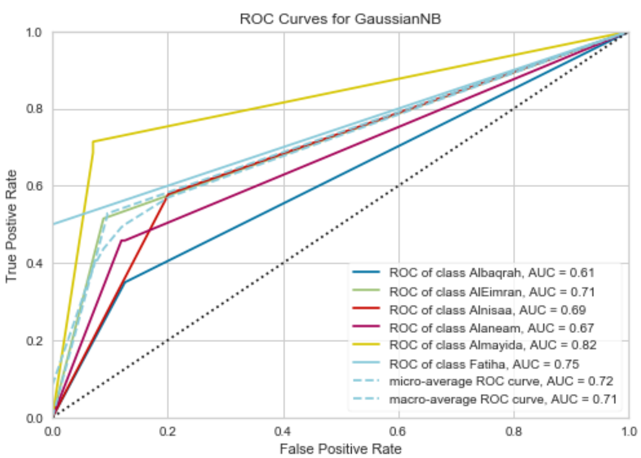


Fig. 13. ROC Curves – GaussianNB.

Finally, SVC [3] and GaussianNB classifiers were implemented for each verse of each Surat and measured the results in terms of the area under the curve (AUC) (see Fig. 14 and Fig. 15) [8]. The experimental results have shown that the proposed model had significant impacts on the multi-class Holy-Quran verse classification (see Fig. 16-19).

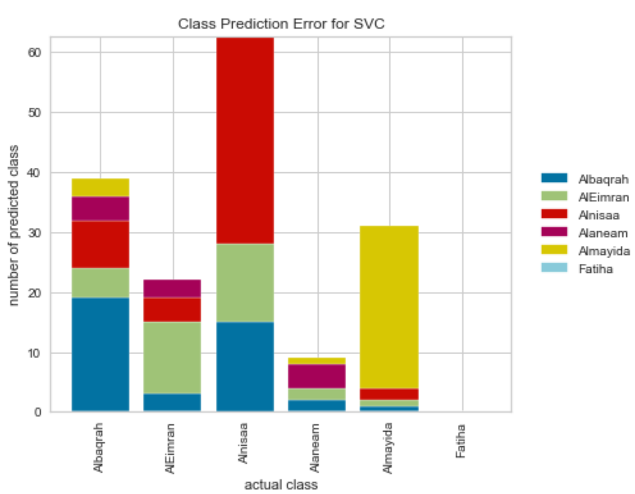


Fig. 14. Class Prediction Error for SVC.

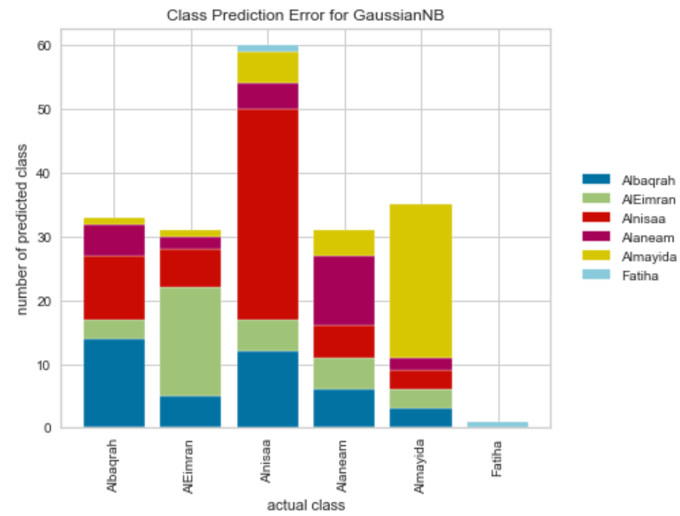


Fig. 15. Class Prediction error for GaussianNB.

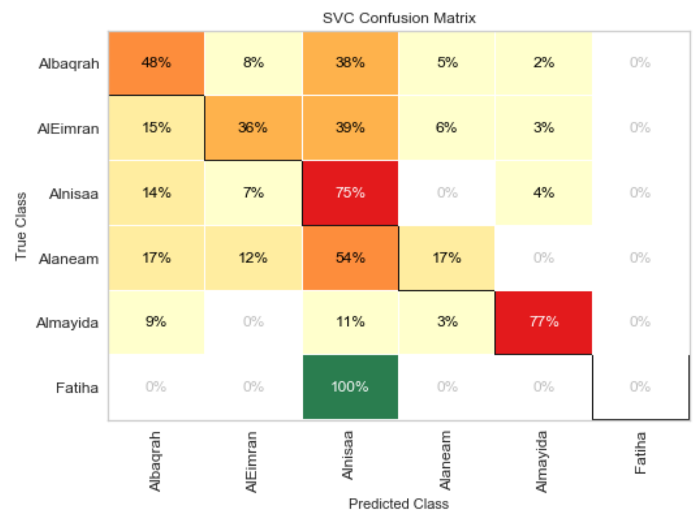


Fig. 16. SVC Confusion Matrix.

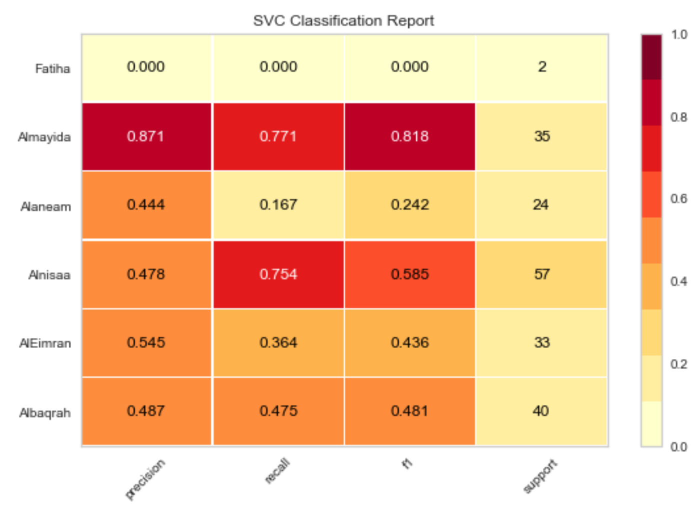


Fig. 17. SVC Classification Report.

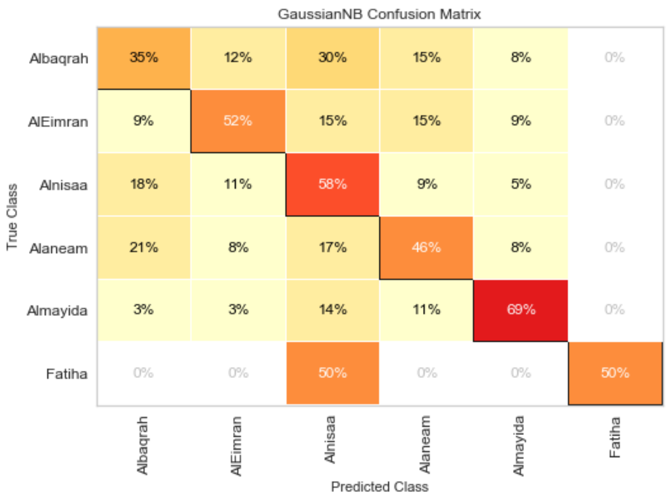


Fig. 18. Confusion Matrix—GaussianNB.

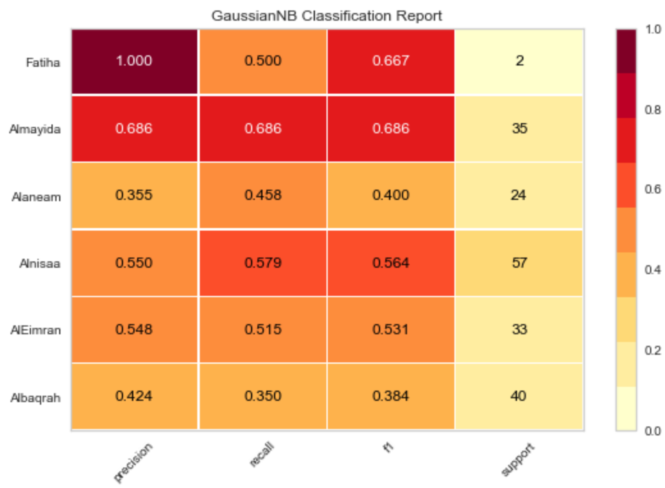


Fig. 19. Classification Report–GaussianNB.

V. Conclusions

Classifying chapters of the Holy Quran is considered a multi-class classification problem. In this paper, the multi-class classification for the Holy Quran corpus was used to train GaussianNB and SVC classifiers to predict the classification of the Quran verses into six surats. Increasing the size of the corpus and improved feature classification may improve the quality and accuracy of the framework. The experiment shows that the SVC provides the best results with an average of 88% f1-score. The research is to be continued by building a larger corpus for the verses of the Holy Quran chapters.

Acknowledgment

This research was fully funded by the Deanship of Scientific Research at Jouf University, AlJouf, Saudi Arabia on 2019/2020 using grant number (40/175); the APC was funded by Jouf University (JU).

References

[1] A. Alsayat and N. Elmitwally, “A comprehensive study for Arabic Sentiment Analysis (Challenges and Applications),” Egypt. Inform. J., p. S1110866519300945, Jun. 2019, doi: 10.1016/j.eij.2019.06.001.

[2] H. M. Abdelaal, A. M. Ahmed, W. Ghribi, and H. A. Youness Alansary, “Knowledge Discovery in the Hadith According to the Reliability and Memory of the Reporters Using Machine Learning Techniques,” IEEE Access, vol. 7, pp. 157741–157755, 2019, doi: 10.1109/ ACCESS.2019.2944118.

[3] A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. Nawi, “Comparative Analysis of Text Classification Algorithms for Automated Labelling of Quranic Verses.,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 7, no. 4, p. 1419, Aug. 2017, doi: 10.18517/ijaseit.7.4.2198.

[4] G. I. Ulumudin, A. Adiwijaya, and M. S. Mubarok, “A multilabel classification on topics of qur’anic verses in English translation using K-Nearest Neighbor method with Weighted TF-IDF,” J. Phys. Conf. Ser., vol. 1192, p. 012026, Mar. 2019, doi: 10.1088/1742-6596/1192/1/012026.

[5] M. N. Al-Kabi, B. M. A. Ata, H. A. Wahsheh, and I. M. Alsmadi, “A Topical Classification of Quranic Arabic Text,” Taibah Univ. Int. Conf. Adv. Inf. Technol. Holy Quran Its Sci., p. 7, 2013.

[6] R. A. Pane, M. S. Mubarok, N. S. Huda, and Adiwijaya, “A Multi-Lable Classification on Topics of Quranic Verses in English Translation Using Multinomial Naive Bayes,” in 2018 6th International Conference on Information and Communication Technology (ICoICT), Bandung, 2018, pp. 481–484, doi: 10.1109/ICoICT.2018.8528777.

[7] D. K. Batubara, M. A. Bijaksana, and Adiwijaya, “On feature augmentation for semantic argument classification of the Quran English translation using support vector machine,” J. Phys. Conf. Ser., vol. 971, p. 012043, Mar. 2018, doi: 10.1088/1742-6596/971/1/012043.

[8] A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. M. Nawi, “A Group-Based Feature Selection Approach to Improve Classification of Holy Quran Verses,” in Recent Advances on Soft Computing and Data Mining, vol. 700, R. Ghazali, M. M. Deris, N. M. Nawi, and J. H. Abawajy, Eds. Cham: Springer International Publishing, 2018, pp. 282–297.

[9] A. Adeleke and N. Samsudin, “A Hybrid Feature Selection Technique for Classification of Group-based Holy Quran Verses,” Int. J. Eng., p. 7.

[10] M. Karan, J. Šnajder, D. Sirinic, and G. Glavaš, “Analysis of Policy Agendas: Lessons Learned from Automatic Topic Classification of Croatian Political Texts,” in Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Berlin, Germany, 2016, pp. 12–21, doi: 10.18653/v1/W16-2102.

Cancelable Face Template Protection using Transform Features for Cyberworld Security

Firdous Kausar
Electrical and Computer Engineering Department
College of Engineering, Sultan Qaboos University
Muscat, Oman

***Abstract*—Cyber world becomes a fundamental and vital component of the physical world with the increase of dependence on internet-connected devices in industry and government organizations. Provision of privacy and security of users during online communication offers unique cybersecurity challenges for industry and government. Intrusion is one of the crucial issues of cybersecurity, which can be overcome by providing the vigorous authentication solutions. Biometrics authentication is used in different cybersecurity systems for user authentication purpose. The cancelable biometric is a solution to rid of privacy problems in traditional biometric systems. This paper purposes a new cancelable face authentication method, which uses Hybrid Gabor PCA (HGPCA) descriptor for cyberworld security. The proposed method uses the wavelet transform for the extraction of the features of the face images by using Gabor filter and Principal Component Analysis (PCA). Later, both types of features have been ensemble by using the simple concatenation scheme. Then scrambling has been applied to the fused features by using the random key generated by the user. So finally, scrambled fused features has been stored in the database which are used for the cancelable biometric authentication as well as recovery. HGPCA achieves "cancelability" and increases the authentication accuracy. The proposed method has been tested on three standard face datasets. Experimental results of the proposed method have been compared with existing methods by using standard quantitative measures that show superiority over existing methods.**

Keywords*—*Cancelable biometrics; face authentication; feature extraction; Gabor filter; Principal Component Analysis (PCA); wavelet transformation

I. Introduction

The number of internet connected devices will reach to 75.44 billion worldwide by 2025 [1] which intensify the risk of cybersecurity breaches such as identity theft, stealing or manipulation of data, credit card frauds, cyber bullying, ransomware and cyberterrorism. Cyber attacks are growing and evolving in prominence every day, causing major damages to industry and government. Most of cyberattacks can be combated by identifying the intruders with proper user authentication. Biometric authentication is considered as one of the most effective user authentication method in cyberworld applications. The biometric can be defined as computerized identification of behavioral or physical uniqueness (e.g. face, gait, fingerprint, voice, iris, etc.) of any person and it must fulfill the standards of universality, uniqueness, collectability, acceptability, and permanence [25]. Currently the existing systems of biometrics (Fig. 1) require further information for some comparisons for generating the templates. The templates generated by biometric systems contain same characteristics across the databases. For example, if some minutiae based system extract the minutia sets of face, these remain similar in different systems.

The biometric data has constant association with specific users, which creates a critical problem. In case of biometric data is compromised from a database by un-authorized users, the original owner of his/her data lose control forever and lose the identity. This makes the biometric templates stand out as a vulnerability of the authentication system because the templates in all databases of related applications have identical characteristics and using similar algorithms. If in one application, a template is compromised, it can be accessed in another application. Further, it is possible, the stored features of templates can be used for the creation of replica demonstration of biometric attributes and can be easily used to access the system. A replica face can be developed from a face template. The compromised biometric template cannot be reissued and discarded. To control this critical problem, the templates are replaced with some biometric features because a person has a specific amount of biometric characteristics. Further, in case the biometric templates are stolen, the attackers can easily attack some other authentication applications which are using the same biometric templates. Due to a person's physical features, the biometric characteristics cannot be effortlessly changed just like keys and passwords [2,3].

The authentication system has vulnerable properties for the biometric templates stored in the databases. The biometric templates in the database have the following risks due to the successful attacks:

1) The attacker can replace the biometric template by an imposter's template for gaining an unauthorized access.

2) The access can be got to the system(s) as per same biometric trait by creating a physical fake from the template available in the database.

3) The attacker can replay the stolen template to the counterpart for gaining the unlawful access.

Therefore, keeping in view the above serious issues, the cancelable biometric has become a prerequisite and has been presented to improve the security of biometric templates. The cancelable biometrics add some additions or alteration to the templates keeping with original biometric data for identification process. If some sets of biometric templates are compromised, these can be easily removed and new biometric templates can be added to the database. The biometric templates can be protected with following properties:

- Template Diversity: It is very necessary that the different templates of a specific user must be used for different applications and should not match each other. The template diversity property should contain the privacy of the data of the user and it shouldn't allow the cross matching between more than one database.
- Template Revocability: The specific template can be easily revoked if it is compromised. In the place of compromised template, the new template can be issued to the user by using the same biometric. It is very important to remember that the new template must not be matched with compromised template. The revocability property mean to cancel the old compromised template, issue the new one template and as well as to cancel the authentication rights of old authenticator associated with compromised template.
- Template Security: The template security property protects the privacy of the biometric data. It must be impossible to get the original biometric template from the secured template. This property should ensure the physical deceiving of biometric and it cannot be theft the template.
- Performance of the System: The system should ensure the identity of the user with high level of confidence. The system should resist the denial attempts. The performance of the system should not be degraded in case of replacement of the templates.

This paper proposed a hybrid rotation invariant features based on random key for face authentication. The scrambling has been applied on hybrid features to secure templates and hybrid rotation invariant features has been selected by applying Gabor and Wavelet Transform based principle component analysis. The paper is structured as follows: in Section 2 (related work), we explained the current developments in the cancelable biometrics field. In Section 3, we applied the Discrete Wavelet Transform on original images for extraction the features by using the Gabor filter and PCA. After extraction of the Gabor and PCA features, we did the fusion of Gabor and PCA features, scrambled the coefficient vectors, and normalized the features and feature matching. In Section 4, there is results and discussion. Finally, the conclusions and perspectives are given in Section 5.

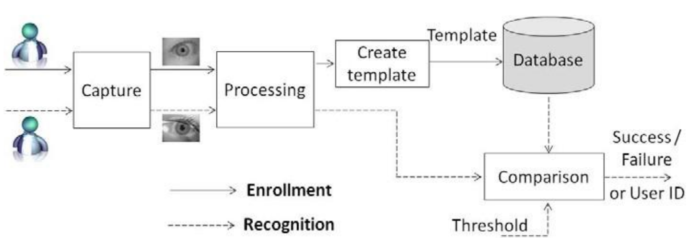


Fig. 1. Architecture of Cancelable Biometric System.

II. Related Work

Cancellable biometric consists in applying an intentional irreversible transformation on the biometric feature, then storing the distorted template in the database. This provides an individual with different biometric templates for different application. Ratha et al. presented the principle of cancelable biometrics in [25]. Lots of solutions can be found in the literature regarding cancelable biometric systems. These systems use some transformation process on the biometric data.

Cancellable multi-biometric approach was investigated in [2]. Multi-biometric is the fusion of different types of biometric information. The main goal of using a multi-biometric is to overcome the limitations such as the non-universality on unimodal biometric systems. Securing distinctive formats of a user independently is not empowering approach in term of security. Therefore, cancellable multi-biometric was designed in a manner to produce a single feature from a mixture of biometric features by applying an irreversible mixture transform. A multi-biometric approach using fuzzy vault scheme was designed and developed in [3] by Nandakumar et al. Their technique was based on two main steps. First, they use an individual template to derive a single multi-biometric template. Securing distinctive formats of a user independently is not empowering approach in term of security. They demonstrated that multi-biometric gives better acknowledgment execution and higher security in contrast with a solitary biometric fault. They have based their experimental tests multi-biometrics vaults on fingerprint and iris.

In [2] Genetic algorithms are applied for multi-biometric cancellable identification. The proposed system was tested via two different biometric face, modalities and voice separately. Use of genetic algorithms shows certain flexibility, since they involved in different part of a biometric authentication system. For more proficiency in their work they connected an upgraded framework in light of the primary combination association to consolidate the diverse data sources. For the transformation methods, we cite BioHashing of Jin et al. [4], in which proposed a two component authenticator in light of the cycle of the inward items between a pseudo-irregular number and the individual unique mark include. The BioHashing provide zero false accept rates.

Connie et al. [5] produce unique palmhash code by applying the hash function on a combination of the palmprint feature and a set of pseudo randomly generated keys. The use of pseudo random keys in the process of generating palmhash code can produce several sets of palmhash codes to be used in multiple applications. It also helps in revocation of biometric keys. Non-invertible property of the palmhash codes make it feasible to be stored in tokens and smartcards for authentication purpose without the threat of retrieving original biometric template from it.

Savvides et al. [6] propose a method to generate cancelable face biometrics by encrypting the biometric template using random convolution kernels. The minimum average correlation energy (MACE) sort relationship channels are utilized to produce the single biometric channel from these convolved encoded formats. The biometric layout can be effortlessly repudiated or re-issued by utilizing diverse arbitrary convolution piece.

Boult et al. [7] use fingerprints template to generate cryptographically secure biotokens. Keeping in mind the end goal to upgrade security in biometric frameworks, biotokens, which they allude to as Biotope™, are embraced to existing acknowledgment plans. Wei-jing et al. [8] presented a strategy to produce particulars dispersion based cancelable unique finger impression layout known as multi-line codes. A multi-line code is created by speaking to the unique mark layout as a number string which is made out of various exceptionally composed minutia codes. The performance of this scheme is ideal with zero percent of equal error rate (EER) as long as secret key remain secure, however accuracy of system deteriorates considerably in the case of a compromise of secret key.

Christian et al. [9] perform score fusion based comparison during the feature alignment process of cancelable iris biometric templates. A biometric framework is upgraded with score level combination by consolidating the correlation scores of numerous comparators. This method also improves the overall accuracy by overcoming the unavoidable degradation in the accuracy of these systems caused by cancelable iris biometric systems.

Marta et al. [10] propose a method to generate non-invertible face templates established on the adaptive bloom-filters. Gabor-based features are extracted from the face image after preprocess it. These extracted features are encoded and binarized. Then binarized features are used to compute Bloom filters which are compared to a reference template in order to calculate the final scores for comparison with a threshold value. Faster authentication is achieved because of the reduced size of templates generated by using bloom filters.

Salman et al. [11] propose a two-factor authentication system consist of password and handwritten signature. The secure biometric templates are generated by applying random projections to signatures templates using secure random keys which are derived from user entered passwords. These irreversible secure biometric templates can be stored in portable devices to provide user authentication for different applications.

Kanagalakshmi et al. [12] presents another method comprising of building distinctive stages, for example, preprocessing, details extraction, post preparing and cancelable and unavoidable layout era. A cancelable format is created utilizing biometrics unique mark highlights. The new technique is tried in view of numerous angles, for example, cancelability, unavoidability and security. Some execution components were utilized also like coordinating time and layout memory utilization. The trial comes about demonstrate that the proposed RMCCP change strategy plays out a decent execution.

Radha et al. [13] introduced the new technique for biometric analysis called BioHashing to generate the cancellable biometrics by using the fingerprints. The advantage of this method is that it does not involve any re-alignment of fingerprints similar to other techniques. The fingerprint translated into pre-defined two dimensional space. Then the Biohashing method is used to achieve the one-way property of the biometric template. This method is highly resistant to any negligible translation error and rotation distortion. The obtained result shows an Equal Error Rate (EER) of less than1%.

Polash et al. [14] presented the cancellable biometrics system, which based on transforming biometric data and feature in order to achieve cancelability. This technique is based on three main steps. The first step consists of performing a twofold random selection of the signals. The got overlap is then anticipated arbitrarily utilizing a projection strategy. Second step comprises of decreasing the element measurement of the arbitrarily anticipated folds utilizing the Principal Components Analysis (PCA). Creating a single template for face biometrics, using K-mean clustering for dimension reduction. Finally, a Linear Discriminant is used to the feature in order to enhance discriminability. The obtained feature is then passed through a classifier to acquire the ultimate authentication performance.

Dwivedip et al. [15] exploit the concept of randomized look-up table mapping for the generation of cancelable iris template. After pre-processing of iris images, the feature vectors are generated in binary matrix form using a 1-D Log Gabor filter. Decimal vectors are generated by using consistent bit vector and perform the decimal encoding. Finally, the cancelable templates are generated by using the lookup table and decimal encoded vector. Experiment results carried on different iris dataset shows that it preserves the transformation properties of the concealable iris templates.

Sree et al. [16] utilized the fluffy vault to create cancellable multimodal biometric layout for face and fingerprints. The details highlight from unique mark are separated by utilizing the crossing number idea and the nearby double example calculation is utilized to extricate the face elements and both are consolidated by highlight level combination. Fluffy vault is made by including copy values and having a mystery key to bolt and open the framework. Execution examination demonstrates that the fluffy vault enhances the execution of the cancelable authentication framework.

Edlira et al. [17] proposed the biometric format assurance conspire by utilizing sprout channels and the idea of engineered layouts, keeping in mind the end goal to delude aggressors called nectar formats. From the input biometric template, feature vector is extracted, then the feature re-arrangement is done by structure-preserving and finally the bloom filter computation is performed on it. The proposed scheme is implemented by conducting experiments on facial authentication. The analysis shows that it achieves the properties of unlinkability, irreversibility and detection of stolen templates of a cancelable biometric system, but it needs to improve templates indistinguishability, pseudonymity, and unobservability. Rathgeb et al. [18,19] propose cancelable biometric template protection schemes based on bloom filter. Binary Feature vectors are extracted from different Iris authentication systems and then transform is applied based on the generic adaptive Bloom filter. Analysis results show that it provides the properties of irreversibility and unlinkability of biometric templates. Further, a high level of security with acceptable authentication rate can be achieved by using the rotation-invariant Bloom filter-based transform.

Jin et al. [20] combines the notion of biometric cryptosystem and cancellable biometrics, and present a key binding scheme for minutiae-based fingerprint bio-metrics. They exploit the idea of chaffing and winnowing (CWS) [21] to confuse the eavesdropper in order to distinguish between real and bogus data and provide confidentiality without using encryption which actually reduces the overall performance of the system.

Yang et al. [26] propose a multimodal cancelable biometric system based on fingerprint and fingervein by using enhanced partial discrete Fourier transform. They design feature level fusion scheme with three different fusion options and provide revocability and irreversibility. Murakami et.al [27] presents the permutation based indexing search scheme for generating cancelable templates. They applied their scheme on face, finger and finger vein datasets to show its performance. The finger vein based templates outperform in term of accuracy as compared to face and finger based templates. Another approach to non-invertible and revocable cancelable face template generation is presented by Wang et al. [28] in which extracted face features are applied repeatedly with random orthonormal transformation (ROT) to generate the cancelable face template. It does not perform well in case of stolen key scenario.

III. Proposed Method

After Keeping in mind the end goal to accomplish the specified cancelable biometric prerequisites, we proposed a Hybrid Gabor PCA (HGPCA) based feature descriptor for cancelable biometric authentication. Fig. 2 demonstrates a general overview of the proposed cancelable biometric framework. The proposed method has been divided into two different phases known as an enrollment and authentication process.

During the enrollment process, the image of a face is captured and feature extraction process has been performed to extract rotation scale invariant features. For features extraction, we have extracted two different types of features by using Gabor filter and wavelet based PCA features and combined. Then these hybrid features have been scrambled by using random key to protect the face template and normalization has been applied to the scrambled features. These features are stored as templates in the database for authentication. If a specific image is compromised, it can be changed by creating a new template. Finally, these features can be used for the authentication process.

During the authentication process, the same steps are performed on the unknown test face image. In this process, first features extraction is performed, then key based scrambling is performed. From that point onward, by using some matching measures, the features compare with the features stored in the database. Detail of the process has been given below step by step.

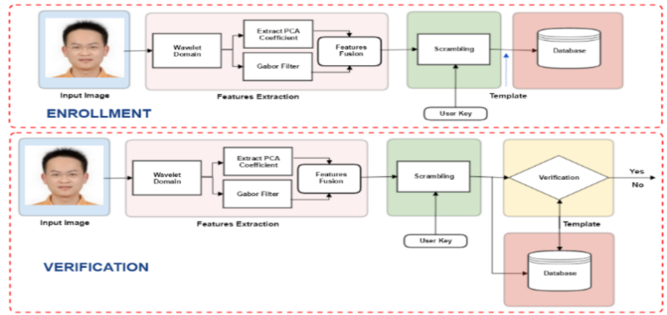


Fig. 2. Block Diagram of Proposed Method.

A. *PCA Feature Extraction using Discrete Wavelet Transform (DWT)*

First the face image is applied with 2D-DWT, which reduces its dimensionality and is helpful in reducing computational overhead. Further, it also provides the multi-resolution data approximation and insensitive feature extraction, which is desirable for security sensitive applications, and transform domain. Then the 2D matrices of face image are converted into 1D image vector for face authentication techniques based on PCA. High-dimensional image vector space is produced by the 1D image vector. The large size and less number of training samples of this high-dimensional image vector space makes the evaluation of covariance matrix more hard.

1) Discrete Wavelet Transform (DWT)): DWT is applied on 1D image vector. The image is passed over a sequence of filter bank stages in order to create its wavelet transform. Afterwards, down sampled is performed in the horizontal direction on these filtered outputs by a factor of 2. Then an identical filtered pair is applied in the vertical direction on each of these signals. We use the symbol of LL, HL, LH and HH to represent the image after decomposition into 4 sub-bands. The unique image characteristics are represented by these sub-bands and are considered as a tinier version of the image. Workflow of DWT is shown in Fig. 3.

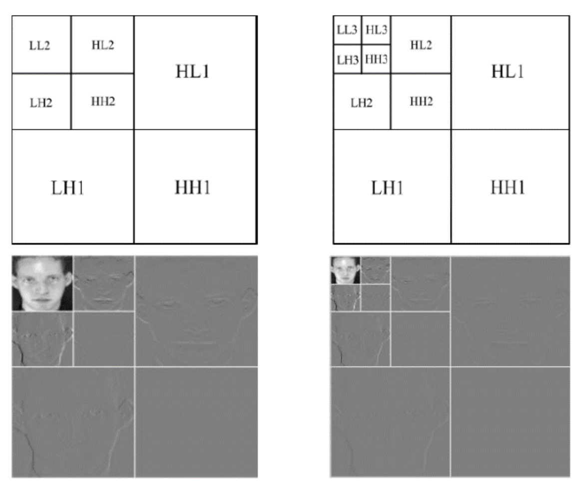


Fig. 3. Wavelet Transform Process.

After applying a 1-level DWT on an image, we get the approximation sub-band LL, the horizontal sub-band LH, the vertical sub-band HL, and the diagonal sub-band HH. Moreover, if we apply a 2-level DWT on the image, we just simply apply another 1-level DWT on the approximation sub-band LL. After applying a 2-level DWT, we also get the approximation sub-band LL2, the horizontal sub-band LH2, the vertical sub-band HL2, and the diagonal sub-band HH2 of the approximation sub-band LL other than sub-bands LH, HL, HH. Applying IDWT to LL, HL, LH, and HH, we get four different frequency's images that are low frequency image, middle-low frequency image, middle high frequency image, high frequency image separately. In this paper, we also apply 2-level DWT so that image dimensions can be reduced. For features extraction, we use the approximation component.

2) Principal Component Analysis (PCA): Although, DWT also reduced the dimensions, but still it has bigger dimensions. We have an image of 256x256 dimensions, then after applying 2D-level, we get an image of 64x64=4096 features that is huge to use for security applications. Therefore, we have used PCA to reduce it further with only significant features. In our proposed enrollment phase, 2-D wavelet is applied at suitable LL subband and then applied 2DPCA to extract most suitable features. During authentication a face image, 2-D wavelet decomposition is applied till specific level and then 2DPCA is applied to extract the features of the face. The Euclidian distance is measured between the mean values of authenticated image and enrollment image in each class to recognize the face.

B. Gabor Filter

We use the Log-Gabor filters (Field, 1987) in our proposed method instead of Gabor (1946) because it is more appropriate for natural image coding as it is more uniform for human visual system quantities. A set of Gabor wavelets is convolved with the input face image. The resulting images are then utilized to extract feature, which produce the salient representation. We illustrate the example of this process in Fig. 4. The original face image shown in Fig. 4(a) is convolved with the Gabor wavelets set of size 4 × 8 shown in Fig. 4(b). Gabor wavelets are used to extract the rotation invariant texture features where we calculated the mean and variance of the Gabor filtered images to find these features. An image feature is used to produce the feature vector.

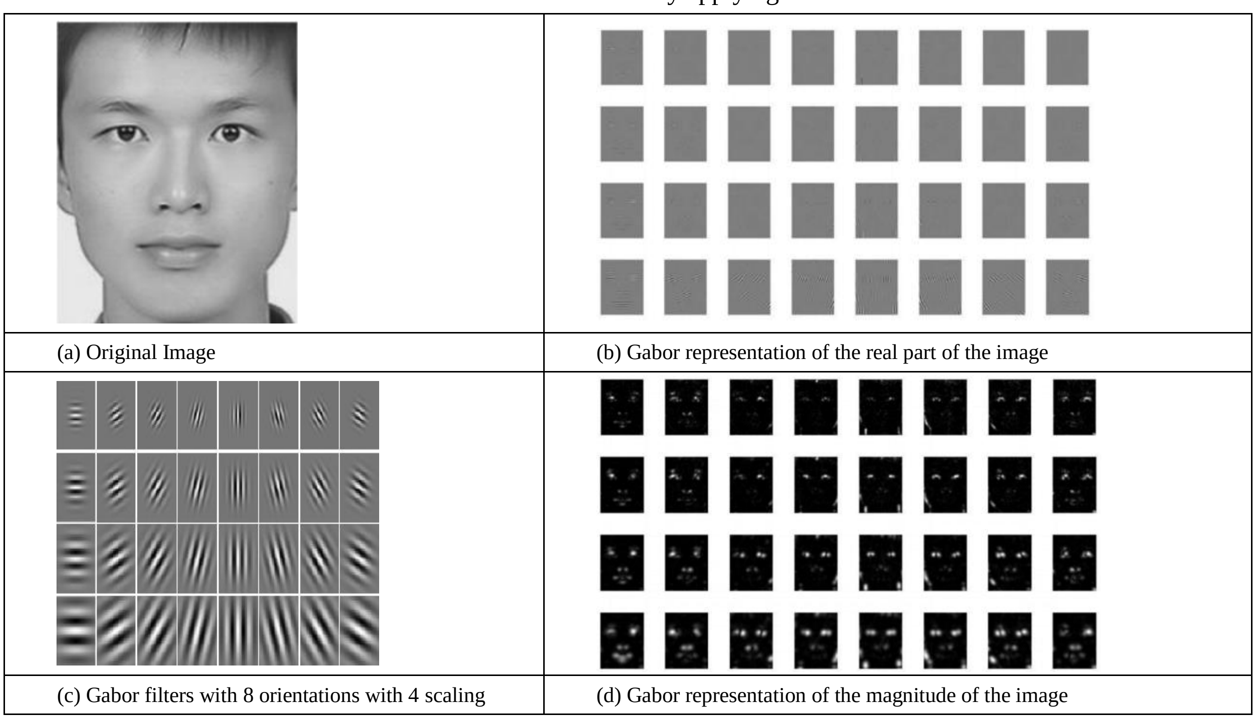
C. Fusion of Gabor and PCA Features

In the previous section, we have extracted Gabor features and wavelet based PCA features. In this section, we have to combine both these types of features. The process of fusion is depicted in Fig. 5. We use a simple process to combine both feature types. We concatenate both types of features and use to store in the database for training and later used for authentication purpose.

GPCA = {Gabor Features, PCA Features}

D. Scrambling the Coefficient Vectors

During the scrambling phase, a random scrambled function depends on the user ID is applied on each weighted coefficient vector. In this scheme, we are using two scrambling functions naming PCA SID and Gabor ID. The weighted normalized PCA coefficient vector p is scrambled by PCA SID while the weighted normalized Gabor coefficient vector ‘i’ is scrambled by applying Gabor ID.



(a) Original Image

(b) Gabor representation of the real part of the image

(c) Gabor filters with 8 orientations with 4 scaling

(d) Gabor representation of the magnitude of the image

Fig. 4. Gabor Wavelet Representation of a Face Image.

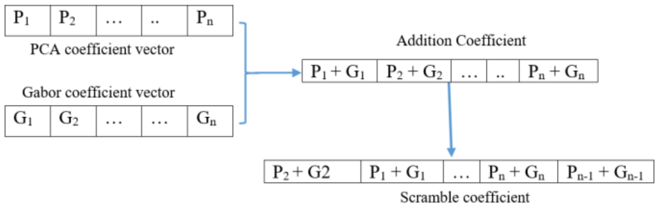


Fig. 5. Fusion of Gabor and PCA Features.

E. Normalization Features

To normalize the features, we use min-max normalization process that is applied on scrambling features. Following method has been used to normalize the features for matching

$$norm_{data} = \frac{(bla - \min(bla))}{\max(bla) - \min(bla)} \quad (1)$$

F. Incorporating the Key Information to the Templates

In our research, we use the 4-digit key, which has been obtained from the user as a password. It is very critical to incorporate key data into the biometric templates. We have used the statistic that the key information can be reflected by ring information, which is non-revocable. The feature descriptor converts to 65 length vector (64 bins plus the ring information) as per the new feature information. The two digits from 4-digit key is taken as two parameters for orientation and scaling. The orientation covers the large variety of conditions with different facial expressions. The scaling covers the different sizes of the face because the practical implementation of face detection system, the system must be able to recognize the face in different sizes. Further, for the enhancement of security, the feature vectors are different for each template of every individual.

G. Feature Matching

During the enrollment process, all features of the faces are stored in the database as a template. During authentication, features of the test image are calculated. The test image is compared with the feature template and calculates a score.

In the coordinating procedure, there could be four conceivable situations: the two patterns for coordinating could be from:

1) The two templates match perfectly in case same user with the same key.

2) The matching of two templates produces high distance in case of the same user with different keys because of different transformations. As a result, these two templates would not be matched.

3) In the third case where we have different users with same key produces high distance because of Gabor Descriptors are different for different faces.

4) In the fourth case of different users with different keys, the transformations would be different and produces a high distance, which considerably reduce the false acceptance rate.

IV. Results and Discussion

A. Dataset

In this paper, the AR, Yale, and FERET Face databases [22, 23, 24], have been used for the evaluation of authentication performance as shown in Fig. 6.

AR database [22] has more than 3200 images and it has all frontal view of 126 subjects, which has captured in the large variety of conditions with different facial expression e.g. smile, anger, neutral expression, right-light on, left-light on, all lights on, scream, wearing scarf, wearing scarf and left-light on, wearing scarf, wearing scarf and right-light on, wearing sun glasses, wearing sun-glasses and left-light on, wearing sun-glasses and right-light on. In this database, the image consists of “81 x 81” array of pixels and second session (same conditions as all expressions). The number of pictures of per person is 26, which have been recorded over the span of two weeks in two different sessions.

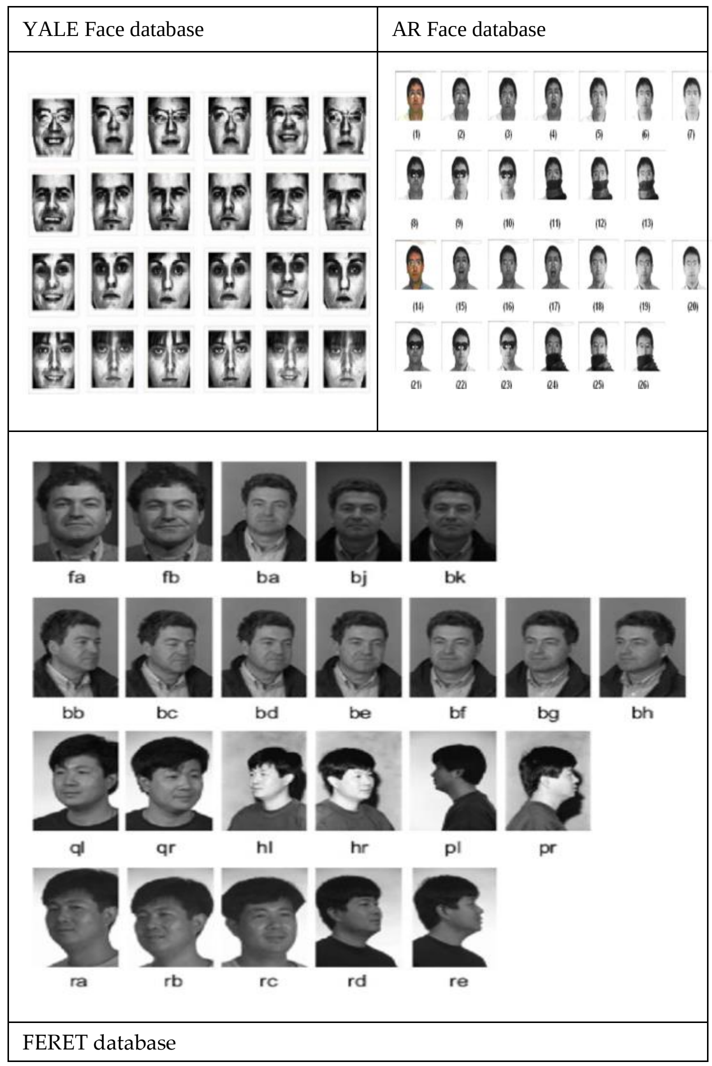


Fig. 6. Sample Images from different Datasets.

Yale database [23] consists of 165 images of 15 persons in grayscale and GIF format. Each person or subject has 11 images one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink.

The FERET [24] database consists of total 14,126 images of 1199 persons with 1564 sets of images and 365 duplicate sets of images.

B. Performance Evaluation

The performance of our proposed method is assessed by measuring the two important parameters including the false acceptance rate (FAR) and false rejection rate (FRR). False acceptance rate is defined as the measure of percentage of accepting the unauthorized user as an authenticated user. On the other hand, the false rejection rate is a measure of percentage of not authenticating the legitimate user. The two parameters together are used to measure the error rate of the system and play an important role to measure the overall accuracy in biometric systems. When FAR and FRR are equal, then it is referred as the equal error rate (EER). We have to minimize this EER in order to achieve high accuracy of the system.

For our experiment, we use the database comprises of 20 samples of each subject. The process consists of two phases i.e. enrollment and authentication. For enrollment phase, 16 images of each subject are used to create the enrollment set while in the authentication phase remaining 4 images of each subject is used to create the authentication set. Authentication process is 1:1 comparison of biometric templates in database as compare to identification, which is 1:n comparisons and requires more time and effort. We performed the authentication instead of identification for our experiments. We perform the comparison of each of the 4 images in the authentication set with the 16 corresponding images in the enrollment set to calculate the false-rejection. Overall, it computes total of 64 comparisons for calculation of FRR of each subject. Thus, for 10 subjects the total number of comparisons is 640.

The false rejection rate is computed by using the following equation:

$$FRR = \left(\frac{no\ of\ false\ reject}{number\ of\ comparisons}\right) x100\% \quad (1)$$

Next, we approximate the mean FRR after computing the FRR for all the classes. We calculate the FRR for all the subjects on for three databases and found the average values for Yale 0.3, AR 0.671 and FERET 0.753. The number of false rejection and the false rejection rate for each subject class on different data set is shown in Table I.

We compute the false acceptance by perform the 1:1 comparison among each enrolled subject with the rest of 9 authentication subjects set. For each subject, it requires to perform total of 576 (9 x64) comparisons and for overall 10 subject it would require 5760 comparisons.

FAR for each subject is computed by using the following expression.

$$FAR = \left(\frac{no\ of\ false\ accept}{number\ of\ comparisons}\right) x100\% \quad (3)$$

TABLE. I. FRR RESULTS ON DIFFERENT DATASETS

	YALE Database		**AR Face database**		**FERET Database**	
Subject#	**False Reject**	**False Reject Rate %**	**False Reject**	**False Reject Rate %**	**False Reject**	**False Reject Rate %**
0001	0	0	1	1	0	0
0002	0	0	0	0	0	0
0003	0	0	0	0	2	1.45
0004	0	0	0	0	0	0
0005	0	0	2	1.32	3	1.89
0006	0	0	0	0	1	1
0007	1	1	3	1.87	2	1.3
0008	0	0	2	1.52	0	0
0009	1	1	1	1	3	1.89
0010	1	1	0	0	0	0

Experiments results are shown in Table II. We found that FAR is 0.52%, 0.61%, and 0.58% for AR, YALE, FERT database respectively. Similarly FRR is 0.89%, 1.52%, and 1.23% for AR, YALE, FERT database respectively. It is possible to achieve slightly different results if the input images are of low quality because quality of image plays an important role in the accuracy of biometric systems. Number of test subjects may also produce variation in achieved results. Further, the security of system can be enhanced by adding more biometric feature dimensions but it decreases the authentication accuracy.

Table III shows the comparison of EER of our proposed method with some other face based cancelable biometric schemes. Average EER of our proposed method for all three databases is 0.39 which is better than other methods [10][28].

C. Revocability Analysis

The revocability is most imperative property of the cancelable biometrics scheme. The revocability means to cancel the old compromised template, issue the new one template and as well as to cancel the authentication rights of old authenticator associated with compromised template. In any case, the relationship among the templates issued from a similar user ought to be kept negligible so that the invader can barely take in any helpful data from numerous templates. We utilized the normalized mutual information (NMI) to show the relationship between two templates explained a

$$NMI(X;Y) = \sum_x \sum_y P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right) \quad (4)$$

where P(x, y) is the joint probability of x and y, and P(x) and P(y) are the marginal probabilities of x and y respectively.

Table IV shows the NMI values of fixed-length binary MLC for all tested datasets. In average for all three datasets, the NMI value is approximately 0.114, meaning that two templates generated from the same fingerprint using different random keys share 11% of mutual information and are slightly correlated.

TABLE. II. RESULTS ON DIFFERENT DATABASES

Database	FRR	FAR	EER%	GAR at FAR 0.1	GAR at FAR 0.01
AR	0.89	0.52	0.36	97.782	96.253
YALE	1.52	0.61	0.24	98.328	97.185
FERET	1.23	0.58	0.58	96.316	94.371

TABLE. III. EQUAL ERROR RATE (EER) OF DIFFERENT CANCELABLE BIOMETRICS

Cancelable biometric schemes	Equal error rate (%)
Bin Scheme II [10]	5.41
PCA based ROT-O [28]	2.09
Proposed Method	0.39

TABLE. IV. REVOCABILITY OF FIXED-LENGTH BINARY MLC TEMPLATE MEASURED IN NMI

Dataset	NMI
YALE	0.061
AR	0.120
FERET	0.162

D. Real-Valued Templates

The execution of genuine key case for all datasets is perfect (zero EER). Though, on account of compromised random keys, the execution decays of course. Table IV reports the real valued templates under stolen-key situation. We used both Dice's likeness and inner product to watch the pattern of execution for various comparability measures. Table V shows that proposed method achieve good dice coefficient as well as inner product.

E. Non-Invertible Transformation

The transformation process is a non-invertible spatial transformation consisting of a random re-mapping of the 72 features to shuffle the original location. Therefore, theoretically 72! (big value) different transformations can be obtained. We have used random based scrambling that also shuffle the original features. In this way, even the features are not in original form, features are shuffled randomly so that attacked should not know the original position of the features. Random seed is used by the key that has been taken from the user. So, at different level security has been added to secure the template. The transformation procedure is a non-invertible spatial change comprising of an irregular re-mapping of the components to rearrange the first location. In our research, we scrambled the features of the images for changing the original position of the features. We have taken one digit from 4-digit key as a seed point to generate pseudorandom number for the purpose of security and this process scramble the positions of the features. This non-invertible spatial transformation process ensures the security of the system.

TABLE. V. PERFORMANCE OF THE PROPOSED METHOD IN STOLEN-KEY SCENARIO OVER DIFFERENT DATASETS

Similarity measure	YALE	AR	FERET
Dice's coefficient	3.417	3.772	4.352
Inner Product	2.942	3.214	4.014

F. Perform Face Authentication using the Key Incorporated Templates

To match two feature point maps, which we taken from 4-digit key from the users, for the purpose of orientation and scaling, the average of the gap scores between all covering highlight focuses is computed and utilized as the coordinating score between two element point maps as shown in Table VI. Further, we altered the Euclidean distance based coordinating calculation by mulling over the transformation. The exceptional transformation mapping is the same for both enrolled and test image from a specific user by his/her same key. We check the feature point area data before computing the Euclidean distance. We see the two formats as being from various users and proceed onward to the following examination, if some chunks are observed to be too far away. An assailant can't recuperate the first change with just the radius location information. Thus, we added orientation and scaling information, which can accept correct acceptance and can reduce the matching time with keeping the security of the templates.

Table VII shows the results after effecting rotation and scaling factors. Like we have experimented by rotating the images at different angles and find the accuracy and then test the performance after making scale factor as well and check the accuracy results. These results show that proposed method perform well even in rotation as well scaling effect. Occasionally, we have to take pictures that are rotated or at different angles. So, we also need to check whether proposed method works well in these cases or not. But we have extracted rotation and scale invariant features therefore proposed method works in these cases well.

TABLE. VI. AUTHENTICATION PERFORMANCE OF THE PROPOSED METHOD

Dataset	PCA + Gabor	PCA	Gabor
YALE	95.6	92.24	93.42
AR	94.25	81.58	91.23
FERET	91.23	84.19	87.56

TABLE. VII. EFFECT OF ROTATION AND SCALING ON AUTHENTICATION USING THE KEY INCORPORATED TEMPLATES

Dataset	Accuracy (Rotation)	Accuracy (Scaling)
YALE	93.152	93.312
AR	91.252	91.143
FERET	90.051	91.352

V. CONCLUSIONS

The proposed method has aims to introduce face authentication cancelable biometric system with high rate of privacy and accuracy. The cancelable biometric is solution to ride of privacy problems in biometric systems as well as a desirable solution to ensure the revocability and alternative of face template when compromised. The traditional cancelable biometrics approaches often compromise the authentication accuracy. In this paper, we propose a new cancelable face authentication method using Hybrid Gabor PCA (HGPCA) descriptor, which can not only achieve "cancelability", but also increase the authentication accuracy. We apply wavelet transform on original images till the second level for extraction the features via a Gabor filter and PCA coefficient vectors and in next step scramble the same. The produced templates are restored in the database and can be recognized/authenticate with an extraordinary rate of precision. Further, if some restored template is compromised, it can be easily cancelled and replaced. As per the results of our proposed method, the authentication accuracy and privacy are improved as compared to the traditional schemes. Consequently, our proposed method satisfies all conditions of an ideal cancelable biometrics system keeping with high rate of accuracy and security.

As a future work, we would like to apply our proposed method on multi model biometric systems to find out its implication for different cyberworld security applications.

References

[1] Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025 (in billions), https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/ : accessed 25 July 2019.

[2] Canuto AMP, Pintro F, and MC Fairhurst, "Genetic Algorithm and Ensemble Systems for Multi-biometric Cancellable Authentication," Journal of Biometrics and Its Applications, vol. 1, no. 1, 2015.

[3] K. Nandakumar and A. K. Jain, "Multi biometric Template Security Using Fuzzy Vault," in 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, 2008, pp. 1-6.

[4] Andrew Teoh Beng Jina, David Ngo Chek Linga, and Alwyn Gohb, "Biohashing: two factor authentication featuring fingerprint data and tokenised random number," vol. 37, no. 11, pp. 2245–2255, 2004.

[5] T Connie, A Teoh, M Goh, and D Ngo, "Palmhashing: A Novel Approach for Cancelable Biometrics.," Information Processing Letters, pp. 1-5, 2005.

[6] M Savvides, B Kumar, and P Khosla, "Cancelable biometric filters for face authentication.," in 17th International Conference on Pattern Authentication, 2004.

[7] T Boult, W Scheirer, and R Woodworth, "Revocable fingerprint biotokens: Accuracy and security analysis.," in IEEE Computer Society Conference on Computer Vision and Pattern Authentication, 2007.

[8] W Wei-jing, DW Mou-ling, and K Yau-hee, "Multi-line code: A low complexity revocable fingerprint template for cancelable biometrics.," Journal of Central South University, vol. 20, no. 5, 2013.

[9] R Christian R and B Christoph B, "Comparison Score Fusion Towards an Optimal Alignment for Enhancing Cancelable Iris Biometrics," in Fourth International Conference on Emerging Security Technologies., 2013.

[10] M. Gomez-Barrero, C. Rathgeb, J. Galbally, J. Fierrez and C. Busch, "Protected Facial Biometric Templates Based on Local Gabor Patterns and Adaptive Bloom Filters," 2014 22nd International Conference on Pattern Recognition, Stockholm, 2014, pp. 4483-4488.

[11] H Salman, M. A Akbar M. A, S Farrukh, F Mudassar F, and K Zeashan , "Secure biometric template generation for multi-factor authentication, Pattern Authentication," Pattern Authentication, vol. 48, no. 2, pp. 458–472, 2015.

[12] K. Kanagalakshmi & Dr. E. Chandra" A Novel Technique for Cancelable and Irrevocable Biometric Template Generation for Fingerprints", Global Journal of Computer Science and Technology Graphics & Vision. Volume 13 Issue 6 Version 1.0 2013.

[13] N.Radha, S.Karthikeyan "An Evaluation of Fingerprint Security Using Noninvertible Biohash", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.4, July 2011.

[14] Padma Polash Paul and Marina Gavrilova" Cancelable Biometrics: Securing Biometric Face Template", International Journal on Artificial Intelligence Tools 4(1), pp. 25-34, June 2012.

[15] Rudresh Dwivedi, Somnath Dey, Ramveer Singh, Aditya Prasad, "A privacy-preserving cancelable iris template generation scheme using decimal encoding and look-up table mapping", Computers & Security, Elsevier, Available online 24 October 2016.

[16] R. Soruba Sree and N. Radha, "Cancellable multimodal biometric user authentication system with fuzzy vault", 2016 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, pp. 1-6, 2016.

[17] Martiri Edlira, Gomez-Barrero Marta, Yang Bian, Busch Christoph, "Biometric template protection based on Bloom filters and honey template", IET Biometrics, 2016.

[18] C. Rathgeb, F. Breitinger, C. Busch, and H. Baier, "On the application of bloom filters to iris biometrics," IET J. Biometrics, vol. 3, no. 4, pp. 207–218, 2014.

[19] Christian Rathgeb and Christoph Busch, "Cancelable multi-biometrics: Mixing iris-codes based on adaptive bloom filters", Computers & Security, Volume 42, Pages 1–12, May 2014.

[20] Zhe Jin, Andrew Beng Jin Teoh, Bok-Min Goi, Yong-Haur Tay, "Biometric cryptosystems: A new biometric key binding and its implementation for fingerprint minutiae-based representation", Pattern Authentication, Volume 56, Pages 50-62, August 2016.

[21] R.L. Rivest, Chaffing and Winnowing, "Confidentiality without Encryption", MIT Lab for Computer Science, Massachusetts, Cambridge, USA, 1998. http://people.csail.mit.edu/rivest/pubs/Riv98a.pdf

[22] A.M. Martinez and R. Benavente. The AR Face Database. CVC Technical Report #24, June 1998.

[23] Yale face database: cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html

[24] K. Delac, M. Grgic, S. Grgic, Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set, International Journal of Imaging Systems and Technology, Vol. 15, Issue 5, pp. 252-260.

[25] J. Connell, and R.M. Bolle N. Ratha, "Enhancing security and privacy in biometrics biased" IBM systems, pp. 614 - 634, 2001.

[26] W. Yang, S. Wang, J. Hu, G. Zheng, and C. Valli. , "A fingerprint and finger-vein based cancelable multi-biometric system" in Pattern Recognition, vol. 78, pp.242-251, June 2018,

[27] T. Murakami, R. Fujita, T. Ohki, Y. Kaga, M. Fujio and K. Takahashi, "Cancelable Permutation-Based Indexing for Secure and Efficient Biometric Identification," in *IEEE Access*, vol. 7, pp. 45563-45582, 2019.

[28] Wang Y, Plataniotis K: Face based biometric authentication with changeable and privacy preservable templates. Proc of the IEEE Biometrics Symposium 2007, 11-13.

Development of Flipbook using Web Learning to Improve Logical Thinking Ability in Logic Gate

Rizki Noor Prasetyono[1], Rito Cipta Sigitta Hariyono[2]
Faculty of Science and Technology
University of Peradaban
Brebes, Jawa Tengah
Indonesia

***Abstract*—The multimedia-based learning process has great potential to change the way of learning. One of them is a growing multimedia Flipbook from textbooks, with ease of reading and learning without carrying a thick book. The purpose of this study is to produce a Flipbook-assisted Web Learning development product and to improve the ability to logical thinking by using those products. This type of research using the 4D Models model consisting of define, design, develop, and disseminate. The results of the development with expert data validation analysis obtained an overall percentage of 83.92% included in the excellent category. Analysis of the validation of media expert's overall percentage obtained 80% included in the excellent category. And the validation analysis of peers the overall percentage found 84.78% included in the excellent category. From all analysis shows that flipbook-assisted teaching materials web learning is appropriate to be used. The results of the t-test calculations show 10.25 higher than 2.045, the use of flipbook logic gates assisted with web learning increases the ability of logical thinking. And the analysis of N-Gain 0.39 in including the medium criteria, the flipbook gate logic assisted with web learning to increase the ability to logical thinking.**

Keywords—Multimedia based learning; flipbook; web learning; logical thinking; logic gates

I. Introduction

The development of information and communication technology is a concern in all countries. The internet has penetrated the community, even education is outdated if it is not integrated with the internet. The Indonesian government supports through the Minister of Education and Culture Regulation No. 24 of 2012 which is about distance education (PJJ) organized by tertiary education. And approved by Law No. 12 of 2012 concerning higher education specifically in article 31 paragraph 2, approving the implementation of distance education to collect and examine arrangements in education and learning in Higher Education. This shows the importance of online-based learning in discussion and learning in higher education. With online-based learning, students can plan learning time and complete material better.

The multimedia-based learning process has great potential to change the way of learning. Multimedia modules or textbooks become a guide in learning, in accordance with the research of Abu Bakar et al the modules which have visual and graphic elements are very suitable in learning and have the feasibility to be used [1]. One of them is Flipbook, multimedia that developed from textbooks with ease of reading and learning without carrying thick books. In Sugianto's research [2] the virtual module based on Flipbook has an impact on improving learning outcomes, especially in learning digital basics. Arai's research also mentioned that e-learning is capable and effective in learning, especially learning complex formulas [3]. This shows that the use of media is important in learning based on some of this research.

Logic gate becomes material that must be understood by informatics engineering students. Where ordinary algebraic understanding cannot be applied in logic gate material. It takes time and understanding to be accepted by logic. According to Halim [4], the basic material for logic gates is needed by media which helps to facilitate further understanding. As in the study of Ash-Shiddieqy et al stated that the right media can provide effectiveness in improving logical thinking [5]. according to the research, G. Deena and D. K. Raja also state that it takes media such as E-learning because it is a very flexible time and space in learning [6].

From the observation of informatics engineering students at the University of Peradaban 6th semester of the academic year 2017/2018 who have taken digital courses. Students who understand logic gate material is not more than 50% and the difficulty faced is the lack of learning meeting time and teaching materials that still cannot be understood without teaching. This refers to the need for effective learning outside the classroom that is able to provide understanding and time that adjusts from the support of the Indonesian government through the rule of law regarding multimedia and internet-based learning. Refer also from previous research regarding the importance of technology-based learning, both in terms of multimedia and internet-based learning. Reinforced by the results of observations resulted that some students did not understand the logic gate material. This shows that it is needed learning media that are able to provide a flexible alternative in learning that is by developing Flipbook Learning aided Web Learning on logic gate materials to improve the ability of logical thinking.

The limitation of the research is the material that focuses on logic gates because it is related to students' ability to think logically. Another disadvantage is that flipbooks require a computer, smartphone, and internet connection to use. If there is none, then learning/lecture is difficult. Then this study aims only to find out the increase in the use of logic gate material flipbooks to improve logical thinking skills.

II. Literature Review

A. E-Learning

Technological development is a benchmark for change in everything. Industry, society, and even education take part in the effects of technological development. Technology is very influential, one of which is learning media; Indonesian government through Government Regulation No. The year 2005 contains in the organization of education units must be held interactively, inspiring, fun, creative, motivating students, initiatives, channeling talents, interests, supported by physical as well as students who are processed in learning. The explanation provides the view that innovation in developing textbooks is a necessity to support learning.

Electronic Learning (E-Learning) is no longer a common thing in learning innovation under technological developments. There are many types of e-learning that can be used, one of them is E-Book. According to the research Salem et al, e-learning proved beneficial, in that students can gain knowledge or information with a low cost of education and students can learn in all times and places [7]. According to Korat et al that the eBook has the advantage of software-based multimedia that can be used in a proper electronic device [8]. Making E-Books in technological developments now many applications that can be used such as Microsoft Office, Macromedia Flash, Adobe Reader PDF, and Flipbook Maker. Flipbook Maker: The application can be used to make the E-Book look more attractive, simple and easy to use.

Flipbook Maker is a textbook application in the form of files to make it more interesting. According to Hidayatullah Flipbook Maker helps students read directly to feel like opening a book with interesting animation effects [9]. Mulyadi & Wahyuni also stated that the use of Flipbook in the medium of learning to improve students' creative thinking [10]. Also supported by Biological research is the use of Flipbook learning media that can also improve student learning outcomes in physics [11]. So learning integrated with technology through Flipbook provides benefits for students in the learning process.

B. Logical Thinking

Scientific thinking in learning in higher education becomes a necessity in terms of knowledge and skills in order to be growing. Logical thinking is one of the scientific thoughts that must be possessed. Logical thinking cannot be separated from reality, where thinking is the result of the product of reality in the form of the laws of reality which become rules of thinking [12]. For example, in thinking someone adding up the numbers 1 and 1 will produce the number 2, then this is supported by the reality and reality of the facts that can be proven. Heryadi states etymologically logic is a science of reasoning or reasoning that makes the law of thought acceptable [13]. So it can be concluded logical thinking rules of thought that can be accepted by reason and proven by facts.

According to Suryabrata, Logical thinking, in essence, goes through three stages [14]. First the construction of understanding, human thought is based on understanding a concept and then building the main contents in thinking. Second, opinion construction arises where the understanding that has been formed is supported by connecting the existing theoretical concepts and observing objects or events so that it becomes a logical process. And lastly, a Construction conclusion is drawn from an opinion that evaluated the existing concept of knowledge into new knowledge. This process is taken into consideration or decision making to be the correct reasoning process.

Logical thinking is very useful in achieving the learning objectives that have been made. As in Yani et al., all research [15] developments in logical thinking play a role in the achievement of learning objectives, especially in science experiments. Even in research Fadiana et al used a logical thinking ability test to find out the students' initial ability to make learning plans [16]. Supported by Othman et al [17] that logical thinking can be used to analyze the way of learning of students who have low cognitive levels combined with students who think high cognitive.. Some of these studies show the importance of increasing students to improve the ability to logical thinking in their application.

III. Method

A. Research Methods

This research was conducted to create learning media with a flipbook. Then validated by material experts, media experts, and peer reviews. The product developed in the form of a flipbook was then tested on Peradaban University students to find an increase in logical thinking skills. To develop a product used by the 4D model development research. Where the 4D Model process consists of define, design, develop, and disseminate.

B. Define

This stage of the process is carried out planning needs in developing products. With the stages of the study of literature from several reference theories in accordance with product development. Field studies are observing students about the difficulties encountered in lectures.

C. Design

This stage goes through several processes consisting of:

- Looking for material reference: The determination of references for the material is needed to provide guidelines or references in the basic theory and context of the product to be developed.
- Preparation of teaching material: This process is carried out in the preparation of the contents of the development product based on several references.
- Preparation of electronic-based textbooks: This process is an advanced stage in the form of a rough form of electronic textbooks, determining the teaching material to be developed is able to provide an interest
- Design: This process emerges the initial product design which will be carried out to the development stage and validated

D. Development

The development phase is carried out in several stages, namely as follows:

- Development Stage: the draft that has been made then drafted the book product in a format suitable for the device to be used in the form. Flipbooks compiled in the making of designs are made to suit the use of Web Learning. The contents of the material are arranged to develop logical thinking skills.
- Expert Validation Stage: the aim of expert validation is to produce revised product drafts based on expert input. After the draft-1 is validated, the next step is to revise draft-1 so that it produces draft-2.
- Small-scale Trial: draft-2 revised draft-1 then a small-scale trial is conducted. For small scale trials that can be done with a Forum Group Discussion (FGD). This trial was conducted to obtain input and suggestions for improvement as well as assessment. The results of this small-scale trial will be used to revise the initial product to produce draft-3.
- Field Trials: draft-3 is then conducted field trials or large-scale trials, the purpose of this stage is to test the feasibility of the product.

E. Disseminate

This stage is the stage of using Web Learning-assisted Flipbook on logic gate material to measure the increase in logical thinking skills.

F. Population and Research Sample

The population for the selection is 6th-semester students (even) with a total of 30 students or one class. The sampling technique used is "saturated sample" meaning that the entire population that has been determined is used as a research sample. This is because the total population of 6th-semester students is relatively smaller.

G. Product Validity Analysis

The method of collecting product validity data uses a questionnaire, containing statements addressed to material experts, media experts, and peer reviews. Analysis of product data using a Likert scale questionnaire with a score of 1 to 4, with the rating category (4) very good, (3) good, (2) Poor, (1) very poor. Material experts observe and validate products on the aspect of content, aspects of material accuracy, stimulate curiosity, presentation and use of web learning on a flipbook. Media experts have validated the aspects of presentation, general appearance, language and readability, and the use of web learning on a flipbook. For peer review covering all aspects because they are practitioners in teaching.

TABLE. I. QUANTITATIVE SCORE RANGE FOR EACH CATEGORY

No.	Quantitative Score Range	Category
1	$\bar{X} > (M + 1.5\ Si)$	Very Good
2	$M < \bar{X} \leq (M + 1.5\ Si)$	Good
3	$(M - 1.5\ Si) < \bar{X} \leq M$	Poor
4	$\bar{X} \leq (M - 1.5\ Si)$	Very Poor

The results of the validation questionnaire are then analyzed with a range of categories according to Table I.

For the value $\bar{X}$ is the average of the scores for each category of validators. M is the median of each category score. then Si is the standard deviation of each score.

Analysis of the ideal percentage of flipbook products using web learning with equation 1.

$$Percentage\ (\%) = \frac{total\ validation\ score}{maximum\ score} \times 100\% \quad (1)$$

H. Analysis of Increased Logical Thinking Ability

The analysis compares the conditions before and after the field test so that the hypothesis is comparative using the t-test for two groups of data from one pair of sample groups as follows:

$$t = \frac{M_d}{\sqrt{\frac{\sum X_d^2}{n(n-1)}}} \quad (2)$$

Hypothesis for t-test from Eq. 1 for two data groups from one paired sample group namely:

H_0 = There is no difference in mean values

H_a = There is a difference in the mean value

Analysis to test the improvement of logical thinking skills is by:

$$gain = \frac{skor\ post\ tes - skor\ pretes}{Skor\ maximum - skor\ pretes} \quad (3)$$

The gain criteria are as follows:

Height : $g \geq 0.7$

Medium : $0.7 > g \geq 0.3$

Low : $g < 0.3$

IV. DATA ANALYSIS

Data analysis performed is product validity and analysis of logical thinking improvement. Data validity analysis of experts using the and analysis of logical thinking improvement Microsoft Excel application, which has been set up for calculating the validity and conclusion of its category.

Analysis of the validity of the material experts from the overall score criteria is in Table II. An analysis of the recap of overall validity by media experts according to Table III then for the overall analysis of data from the peer review according to Table IV.

Validity analysis by material experts containing 14 assessment criteria aimed at practitioners of logic gate material. The questionnaire which contains 14 criteria has 2 aspects of content, 2 aspects of material accuracy, 3 aspects of stimulating curiosity, 2 aspects of presentation, and 5 aspects of the use of web learning. The analysis also shows that each of these aspects has a very good category can be seen in Table VI.

The media experts in accordance with Table III show that they have 18 validity assessment criteria, which refer to learning media practitioners. where each aspect is 5 aspects of presentation, 5 aspects of general appearance, 3 aspects of language and readability, and 5 aspects of the use of web learning.

TABLE. II. ANALYSIS OF SCORE VALIDITY MATERIAL EXPERTS

Overall Validity Score	
Total criteria	14
Maximum score	56
Minimum score	14
M	35
Sb	7
Average Score (*X*)	47

No	**Range Score**			**Category**
1		*X*>	45.5	Very Good
2	35	<*X*≤	45.5	Good
3	24.5	<*X*≤	35	Poor
4		*X*≤	24.5	Very Poor

Conclusion :	
From the results of the analysis of the validation of the material experts, it was found that flipbook assisted by web learning was included in the category	Very Good

TABLE. III. ANALYSIS OF SCORE VALIDITY MEDIA EXPERT

Overall Validity Score	
Total criteria	18
Maximum score	72
Minimum score	18
M	45
Sb	9
Average Score (*X*)	62

No	**Range Score**			**Category**
1		*X*>	58.5	Very Good
2	45	<*X*≤	58.5	Good
3	31.5	<*X*≤	45	Poor
4		*X*≤	31.5	Very Poor

Conclusion :	
From the results of the analysis of the validation of the material experts, it was found that flipbook assisted by web learning was included in the category	Very Good

The validity criteria in the peer review questionnaire are more than the others. The total overall assessment criteria in the peer review were 23 criteria, consisting of 2 aspects of content, 2 aspects of material accuracy, 3 aspects of stimulating curiosity, 5 aspects of presentation, 5 aspects of general appearance, 1 aspect of language and readability, and 5 aspects of the use of web learning.

Data analysis increased logical thinking according to Table V. Data were taken through the pre-test before the treatment and post-test after the use of a flipbook assisted with web learning. Table V shows the analysis of the t-test and N gain from the pretest and posttest data.

TABLE. IV. ANALYSIS OF SCORE VALIDITY PEER REVIEWS

Overall Validity Score	
Total criteria	23
Maximum score	92
Minimum score	23
M	57.5
Sb	12
Average Score (*X*)	78

No	**Range Score**			**Category**
1		*X*>	74.75	Very Good
2	57.5	<*X*≤	74.75	Good
3	40.3	<*X*≤	57.5	Poor
4		*X*≤	40.25	Very Poor

Conclusion :	
From the results of the analysis of the validation of the material experts, it was found that flipbook assisted by web learning was included in the category	Very Good

TABLE. V. DATA ANALYSIS T-TEST AND N GAIN

SAMPLE	**PRETES**	**POSTES**	**GAIN (d)**	**Xd**	**$(Xd)^2$**	**N GAIN**
1	32	64	32	11.2	125	0.47059
2	44	64	20	-0.8	0.64	0.35714
3	64	65	1	-19.8	392	0.02778
4	64	76	12	-8.8	77.4	0.33333
5	52	76	24	3.2	10.2	0.5
6	48	75	27	6.2	38.4	0.51923
7	68	56	-12	-32.8	1076	-0.375
8	48	65	17	-3.8	14.4	0.32692
9	68	78	10	-10.8	117	0.3125
10	64	77	13	-7.8	60.8	0.36111
11	48	75	27	6.2	38.4	0.51923
12	36	56	20	-0.8	0.64	0.3125
13	52	67	15	-5.8	33.6	0.3125
14	68	80	12	-8.8	77.4	0.375
15	64	78	14	-6.8	46.2	0.38889
16	28	56	28	7.2	51.8	0.38889
17	56	76	20	-0.8	0.64	0.45455
18	32	65	33	12.2	149	0.48529
19	28	65	37	16.2	262	0.51389
20	32	68	36	15.2	231	0.52941
21	32	68	36	15.2	231	0.52941
22	56	78	22	1.2	1.44	0.5
23	48	67	19	-1.8	3.24	0.36538
24	56	86	30	9.2	84.6	0.68182
25	48	67	19	-1.8	3.24	0.36538
26	36	67	31	10.2	104	0.48438
27	36	68	32	11.2	125	0.5
28	52	78	26	5.2	27	0.54167
29	64	79	15	-5.8	33.6	0.41667
30	72	80	8	-12.8	164	0.28571
AVERAGE			20.8			0.39281
TOTAL					3581	
Standard deviation (S)					4.12	
	t-COUNT		t-TABLE			
	10.25257438		DK=N-1=29		α=5 %	
			2,045			

V. RESULTS AND DISCUSSION

Media plays a role in the learning process because it can represent abstract material that cannot sometimes require extra explanation [4]. The most common media is teaching the material as a learning guide. Teaching materials developed in the form of logic gate flipbooks have characteristics; there are guidelines for the use of teaching materials that have an explanation in the content of teaching materials. Teaching materials developed in the form of logic gate Flipbooks have characteristics, there are guidelines for the use of teaching materials according to Fig. 1 which has an explanation in the content of teaching materials.

Media including textbooks helps in the implementation of learning, especially in higher education because the learning process has a short time. The aspect of fulfilling the development of teaching materials is very important to get decent and valid results. Appearance, material, language, and technical aspects are included in the assessment so it needs to be assessed by practitioners [18].

In Fig. 2 each sub material has a series that students can try out through the livewire application. Train for psychomotor and cognitive activities to explore each material being taught. Research states that there is an influence in thinking patterns of critical thinking with interactive learning processes in the classroom.

The results of this research were tested for eligibility through product validity testing by material experts, media experts, and peers. The validity of the product is developed by measuring the quality of the product and the percentage of product ideals. The recapitulation of the validation test data analysis from material experts, media experts, and peers in each aspect is presented in the following Table VI.

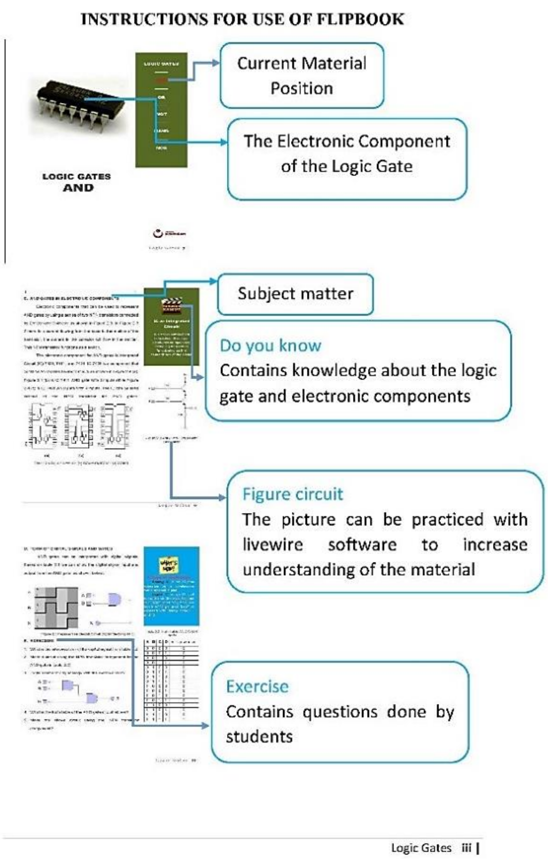


Fig. 1. Instructions for use of Teaching Materials.

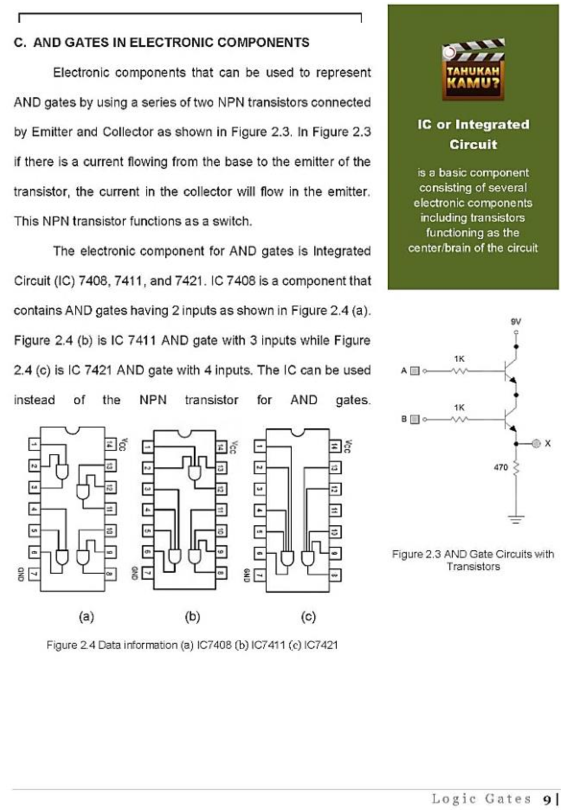


Fig. 2. Content of Flipbook Material.

TABLE. VI. RECAPITULATION OF VALIDATION ANALYSIS OF MATERIAL EXPERTS, MEDIA EXPERTS, AND PEER REVIEW FOR EACH ASPECT

Aspect	Material Expert		Media Expert		Peer Review	
	$\bar{X}$	Category	$\bar{X}$	Category	$\bar{X}$	Category
The content aspect	7	Very good			7	Very good
Material accuracy	7	Very good			7	Very good
Stimulate curiosity	10	Very good			10	Very good
Presentation	7	Very good	16	Good	17	Very good
The use of web learning in a flipbook	16	Good	20	Very good	17	Very good
General appearance			17	Very good	16	Good
Language / Readability			9	Good	4	Very good

Each expert is given a questionnaire that shows aspects according to their expertise and then analyzed to get the validity value of the product being tested. From the results of the data taken from the expert assessment material according to Table II, validation obtained an average score of 47 which according to the attachment of the score is included in the range $\bar{X}$ > 45,5 meaning in the very good category. Data taken from media experts according to Table III the results obtained an overall analysis of product validation obtained an average score of 62 which according to the attachment of the score is included in the range $\overline{X}$ > 58,5 meaning in the very good category [19]. Collecting data by peers reviews according to Table II validation of products obtained an average score is 78 wherein the score is included in the range of $\bar{X}$ > 75 means in the excellent category.

After analysis of the average score of validation data, then analyzed the percentage of ideals of the product. The percentage of product ideals by material experts that is 83.92% is included in the range of the percentage score $\bar{X}$ > 81,25 % in the very good category. Percentage of product ideals by media experts calculated according to overall aspect criteria is 80% included in the range of percentage 62,5% <$\bar{X}$< 81,25 % in the good category. The calculation of the overall percentage of product ideals for peers found 84.78% included in the range of score percentage $\bar{X}$ > 81,25 % in the very good category.

From the calculation results that have been presented both product quality analysis and percentage analysis of ideals, it is found that the flipbook gate logic web-assisted web learning products are in the very good category. The conclusion is that the product is feasible or valid to be tested and used. After product development and validity testing, it is then implemented in small groups. The first product trial was conducted at informatics engineering students for upper semester classes who already knew digital material. The trial was conducted 15 samples which obtained an average score of 81% included in the range $\bar{X}$ > 81.25% meaning that in the good category. Recapitulation of student response data analysis in each aspect is presented in Table VII.

From Table VII shows that student responses from each aspect show that the average product is considered by the sample of users included good category. Then it can be concluded that from the data that the product was received then it could be continued for use to determine the effectiveness of the product. Further implementation to find out the increase and effectiveness of product use. In accordance with the analysis of the data shows that the use of flipbook gate logic assisted by web learning increases the ability of students to think logically by 20.8%. Then from the results of t-test calculations for two groups of data from one paired sample group shows the value of t-count 10.25 which means t-count> t-table, where 10.25> 2.045 then H_0 is rejected. It can be concluded that there is a difference in the average value between before and after using the flipbook gate logic assisted web learning, which shows that after the use of products the ability of logical thinking increases. In line with the research of Divayana et al [20] shows that in the development of flipbooks as a digital book design in both categories as a source of teaching.

Reinforced by the results of the N-GAIN analysis in accordance with Appendix 13, the pretest and posttest values had an N-Gain value of 0.39 which included in the range of 0.7> g ≥ 0.3 moderate criteria. These results indicate that the flipbook gate logic is assisted by web learning to increase the ability of logical thinking to be effective and appropriate for use in the lecture process. Supported by research on the effectiveness of web-based learning support in the learning process because the process is easy and has unlimited access [21].

TABLE. VII. PERCENTAGE OF RESPONSE DATA FOR EACH ASPECT

No	Aspect	Percentage of the ideal score (%)	Category
1	Stimulate curiosity	80,83	Good
2	Presentation	81,66	Very Good
3	The use of web learning in a flipbook	80,41	Good
4	General appearance	80	Good
5	Language / Readability	80,83	Good

VI. CONCLUSION

Based on the results of data analysis and discussions that have been carried out, it can be concluded that the results of the validation of material experts, media experts and peers show the products in the excellent category. Then from the first product trial conducted in informatics engineering students obtained an average score of 81% included in the range $\bar{X}$ > 81.25% meaning in the good category. The results of the t-test calculations show t-count> t-table, where 10.25> 2.045, the use of flipbook gate logic assisted with web learning increases the ability to think logically. And the analysis of N-Gain 0.39 in including medium criteria, then the flipbook gate logic assisted with web learning to increase the ability to think logically effective.

Further research can be developed as follows:

1) Developing more interactive E-learning by involving many interpretations.

2) Looking for influence from the positive and negative impacts of developments in the current digitalization era.

3) Development of software programs for learning.

REFERENCES

[1] M. A. Bakar and M. Mukhtar, “The Development of a Visual Output Approach for Programming via the Application of Cognitive Load Theory and Constructivism,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 11, pp. 305–312, 2019.

[2] D. Sugianto, A. G. Abdullah, S. Elvyanti, and Y. Muladi, “Modul Virtual: Multimedia Flipbook Dasar Teknik Digital,” Innov. Vocat. Technol. Educ., vol. 9, no. 2, pp. 101–116, 2017.

[3] K. Arai, “Comprehensive e-Learning System with Simulation Capabilities for Understanding of Complex Equations,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 11, pp. 330–335, 2019.

[4] A. Halim and M. S. Sumbawati, “Pengembangan Media Pembelajaran Berbasis Adobe Air Kompetensi Dasar Gerbang Dasar Rangkaian Logika Di Smk Negeri 2 Bangkalan,” Pendidik. Tek. Elektro, vol. 05, no. 01, pp. 115–122, 2016.

[5] M. H. Ash-Shiddieqy, A. Suparmi, and W. Sunarno, “The effectiveness of module based on guided inquiry method to improve students’ logical thinking ability,” J. Phys. Conf. Ser., vol. 1006, no. 1, 2018.

[6] G. Deena and D. K. Raja, “Designing an Automated Intelligent e-Learning System to Enhance the Knowledge using Machine Learning Techniques,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 12, pp. 112–119, 2019.

[7] S. Issa and A. Zoubi, “E-learning Benchmarking Adoption : A Case Study of Sur University College,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 11, pp. 463–470, 2019.

[8] O. Korat, T. Graister, and C. Altman, “Contribution of reading an e-book with a dictionary to word learning: Comparison between kindergarteners with and without SLI,” J. Commun. Disord., vol. 79, pp. 90–102, 2019.

[9] M. S. Hidayatullah, "Pengembangan Media Pembelajaran Berbasis Flip Book Maker Pada Mata Pelajaran Elektronika Dasar Di Smk Negeri 1 Sampang," J. Pendidik. Tek. Elektro, vol. 05, pp. 83–88, 2016.

[10] D. U. Mulyadi, S. Wahyuni, and S. Wahyuni, "Pengembangan Mediaflash Flipbook Untuk Meningkatkanketerampilan Berfikir Kreatif Siswa Dalam Pembelajaran Ipa Di Smp," J. Pembelajaran Fis., vol. 4, no. 4, pp. 296–301, 2016.

[11] S. Hayati, A. S. Budi, and E. Handoko, "Pengembangan Media Pembelajaran Flipbook Fisika untuk Meningkatkan Hasil Belajar Peserta Didik," Pros. Semin. Nas. Fis. SNF2015, vol. IV, pp. 49–54, 2015.

[12] J. S. Ab, G. Margono, and W. Rahayu, "The Logical Thinking Ability: Mathematical Disposition and Self-Regulated Learning," J. Phys. Conf. Ser., vol. 1155, no. 1, 2019.

[13] D. Heryadi, "Menumbuhkan karakter akademik academic characters development through logical-based lecturing," J. Pendidik. dan Kebud., vol. 1, no. 3, pp. 317–330, 2016.

[14] S. Suryabrata, "Psikologi Pendidikan," in Psikologi Pendidikan, Jakarta: PT Rajagrafindo Persada, 2012, p. 354.

[15] A. Yani, N. Dewi, and R. Rosita, "Connecting Concepts Test: An Alternative Instrument Form to Improve Students' Logical Thinking Ability in Conceptual Understanding (Case Study on Geography Subject)," IOP Conf. Ser. Earth Environ. Sci., vol. 145, no. 1, 2018.

[16] M. Fadiana, S. M. Amin, A. Lukito, A. Wardhono, and S. Aishah, "Assessment of seventh grade students' capacity of logical thinking," J. Pendidik. IPA Indones., vol. 8, no. 1, pp. 75–80, 2019.

[17] M. Othman, A. F. Rosmani, S. S. M. Fauzi, and U. H. Mazlan, "The Impact of Pair Programming on Students' Logical Thinking: A Case Study on Higher Academic Institution," Soc. Manag. Res., vol. 16, no. June, pp. 85–98, 2019.

[18] A. Agung, B. D. Amin, A. Yani, and A. Swandi, "Pengembangan Bahan Ajar Berbasis E-Learning Mata Kuliah Fisika Dasar pada Jurusan Biologi FMIPA UNM," Indones. J. Educ. Stud., vol. 21, no. 2, pp. 139–147, 2019.

[19] R. N. Prasetyono and E. Trisnawati, "Pengaruh Pembelajaran IPA Berbasis Empat Pilar Pendidikan terhadap Kemampuan Berpikir Kritis," JIPVA (Jurnal Pendidik. IPA Veteran), vol. 2, no. 2, pp. 162–173, 2018.

[20] D. G. H. Divayana, P. W. A. Suyasa, I. P. W. Ariawan, I. W. E. Mahendra, and G. A. D. Sugiharni, "The Design of Digital Book Content for Assessment and Evaluation Courses by Adopting Superitem Concept Based on Kvisoft Flipbook Maker in era of Industry 4.0," J. Phys. Conf. Ser., vol. 1165, no. 1, pp. 0–6, 2019.

[21] M. Barisone et al., "The effectiveness of web-based learning in supporting the development of nursing students' practical skills during clinical placements: A qualitative study," Nurse Educ. Pract., vol. 37, pp. 56–61, 2019.

Automatic Detection and Correction of Blink Artifacts in Single Channel EEG Signals

G Bhaskar N Rao[1]
Department of Commerce MAHE
Manipal University
Kasturba Nagar
Manipal - 576104

Dr. Anumala Vijay Sankar[2]
ECE Dep. Associate Professor
DE & CS Dept of ECE VRSEC
Chalasani Nagar, Kanuru
Vijayawada, AP-520007

Dr. Peri Pinak Pani[3]
GITAM HBS
Rudraram, Patancheru
Sanga Reddy
Dist. TS - 502329

Aneesh Sidhireddy[4]
ECE – IV Year
Roll No: 16BEC0690
VIT University, Vellore, 632 014
Tamil Nadu, India

***Abstract*—Ocular Artifacts (OAs) are inevitable during EEG acquisition and make the signal analysis critical. Detection and correction of these artifacts is a major problem now a day's. In this paper an energy detection method is used to detect the artifacts and performed wavelet thresholding within the researched zones to protect neural data at non blink regions. Various sets of Wavelet Transform (WT) techniques and threshold functions are collated and identification of the optimum combination for OA's separation is indicated in many research areas including Technology & Management. The output of these methods at blink regions is compared interns of various standard metrics using established techniques of Supply Chains. Results of this study demonstrate that the SWT+HT has better in rejecting the artifacts than other methods in this paradigm.**

***Keywords*—Electroencephalogram (EEG); ocular artifacts; wavelet transform; hybrid threshold**

I. Introduction

EEG is a non-invasive technique used to diagnose brain related diseases and disorders. These signals are frequently corrupted by various types of artifacts during acquisition that come from several sources such as blinking of eyes, cornea movements, vibrations of muscle, heart signals can reduce the clinical utility. Among these ocular activities create significant artifacts due to its larger amplitude and makes the analysis critical. Numerous methods are in use to detect the artifacts, but WT techniques are popular due to its easier implementation [1-4]. Krishnaveni et.al [1] has proposed an artifact detection method based on the relative amplitudes of Artifact Rising Edges (ARE) and Artifact Falling Edges (AFE) at Nth decomposition level. Later it is simplified by the process of coefficient of variation. Usually each spike contains three coefficients; from the number of coefficients at each decomposition level recognize the coefficients pertinent to spikes [2, 3]. However detection of artifacts depends on the selection of a parent wavelet function and associated decomposition quotient. WT is a proven methodology and has shown promise of its utility in ocular artifact paradigm and for single channel EEG signals [3-5]. Majmudar et.al [5] compared the discrete and stationary wavelet transforms using various methodologies on SWT and recommended that DWT was superior in artifact correction .Of late Jianbo Gao et.al proposed a Wavelet based hybrid threshold function for denoising nonlinear time signals [6]. This paper proposed an energy detector method for identification of the artifacts and performs adaptive thresholding to the blink regions for effective removal of ocular artifacts. An interpretation of experimental data using queuing theory is adopted for latency and crystallization of decisions [14].

II. Data Acquisition

Raw EEG segments for this work are taken from physionent (www.physionet.org/ physionet/physiobank ATM/eegmmidb) [7]. EEG signals at frontal channels such as F7, F8, Fp1 and Fp2 are taken for analysis, because these electrodes are placed close to the eyes and EEG signals are most likely to be affected by the ocular artifacts. Analysis is done by taking EEG segments of 10 seconds duration each, since EEG epochs smaller than 12 seconds may be considered stationary. The simulations are carried out in the MATLAB environment within the time-frequency domain.

III. Methodology

1) The raw EEG signal is segregated into blink and non-blink regions by energy detector method and perform wavelet denoising to the identified zones.

2) Each blink region is decomposed by WT methods into approximation and detail coefficients up to 8 levels.

3) Levels range from 8 to 4 and function as an inverse WT to reconstruct the refined EEG signal. Fig 1 illustrates the process of artifact removal of raw EEG signals.

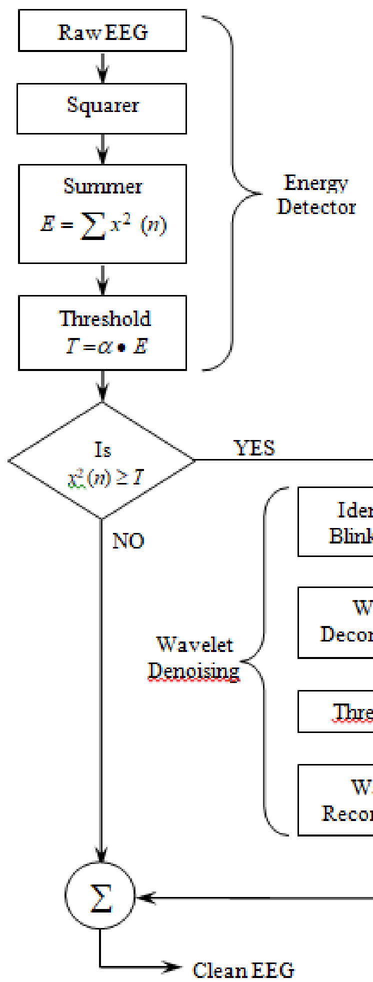


Fig. 1. Flow Chart for Detection and Correction of Blink Artifacts.

IV. Energy Detector Method for Identification of Blink Regions

Correcting the artifacts at blink regions and preserving the neural information at non-blink regions is of very much importance for clinical diagnosis. The EEG signal, tabled evenly between scalp, had amplitude of about 11 µV to 101 µV, whereas artifacts due to ocular activity are 10 to 100 times as that of the EEG signal. The momentous difference in magnitude between the artifacts facilitates to separate the blink and non-blink regions by an energy detector method. Energy detector is a basic signal detection method [8,9]. Blink and non blink regions are segregated by comparing the relative amplitudes of squarer with respective to the threshold level. Threshold level T is estimated from signal statistics as whereas factor α is ranging from 0.001 to 0.01, which decides the accuracy of artifact detection. The detection of multiple artifacts for Fp2 EEG signal is presented below in Fig 2(a) and Fig 2(b).

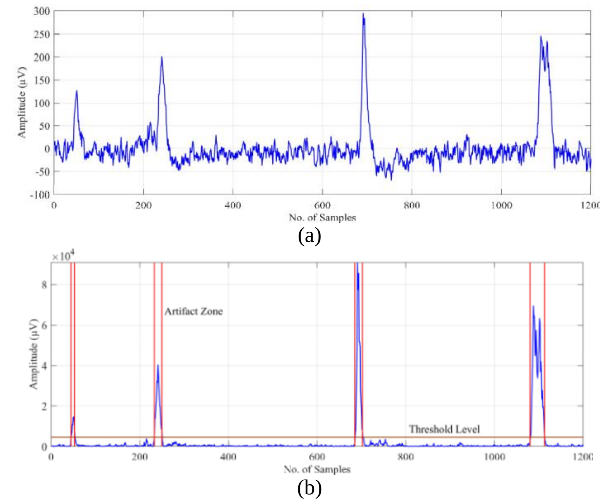


Fig. 2. (a) Fp2 EEG Signal, (b) Multiple Artifacts for Fp2 EEG Signal.

V. Threshold Functions and Performance Metrics

A. Thresholding Functions

In the proposed method, the following thresholds were used for calculating the threshold function and the most optimum one is found.

a) Universal Threshold (UT): UT is a global threshold function, tabulation of values at the Threshold as per Eq. (1).

$$\lambda_i = \sigma_i \sqrt{2 \log N} \quad (1)$$

where N = signal measurement, and is the maximum value at i^{th} decomposition level. σ_i is the average deviation for Wi, which is calculated by the following Eq. (2).

$$O_i = \frac{Median\,|W_i|}{0.6745} \quad (2)$$

where W_i = comprehensive wavelet coefficients at i^{th} level. The numerator is rescaled for a suitable estimator for Gaussian white noise by 0.6834 in the divisor.

b) Statistical Threshold (ST): ST was proposed by Krishnaveni et.al, which is based on the statistics of the signal [5].

The effective statistical threshold is given by

$$\lambda = 1.5 * \mathrm{std}(W_i) \quad (3)$$

in which factor 1.5 is an estimator for standard white Gaussian noise

c) Hybrid Threshold (HT): HT is a combination of UT and ST functions [6], Threshold function at each level is defined by

$$\lambda_i = \mathrm{std}(W_i) * \sqrt{2 \log N} \quad (4)$$

B. Performance Metrics

Working of the threshold functions is validated using power spectral density (PSD), Magnitude Square Coherence (MSC) plots and twin statistical parameters: Artifact Rejection Ratio (ARR) and Correlation coefficient (CC). ARR is the power ratio of the removed artifacts to the clean EEG signals expressed [10].

$$ARR = \frac{\sum_{n=1}^{N}(\mathrm{x[n]} - y\,[n])^2}{\sum_{n=1}^{N} y\,[n])^2} \quad (5)$$

Where x[n] and y[n] represents the contaminated and clean EEG signals. CC is a statistical quantity that shows the degree of similarity or relatedness between two signals expressed as

$$CC = \frac{\sum_{n=1}^{N}(x[n]-\bar{x})(y[n]-\bar{y})}{\sqrt{\sum_{n=1}^{N}(x[n]-\bar{x})^2 \sum_{n=1}^{N}(y[n]-\bar{y})^2}} \quad (6)$$

Based on the power distribution from 0 to 16Hz, SWT+HT has best in rejecting the artifacts, whereas DWT+HT is second

Where x and y represents the mean of raw and clean EEG signals. The Power Spectral Density (PSD) function shows the energy of the signal as a function of frequency. It uses Welch's method (Pwelch). MSC provides the estimate of the frequency coherence between the two signals, which is implemented using 'MScohere' MATLAB function.

VI. Results and Discussions

Denoising of EEG signal is carried by combination wavelet transform methods and threshold functions. Fig. 3 (a) to 3(d) shows the time domain plots of the raw and clean EEG signals using various methods. By visual inspection it is clear that SWT method is superior in correcting the artifacts than DWT method. Threshold function HT is better than other functions in both the methods, whereas ST is the second best. Fig. 4 illustrates the power spectra of DWT and SWT methods using different threshold functions. Threshold functions HT and ST have provided the minimum power at lower frequencies respectively in both the methods [11] and [12]. The MSC plot for FP1 EEG signal is shown in Fig. 5 and Fig. 6, it is observed that the frequency coherence is less at lower frequencies and nearly '1' for all higher frequencies in both the methods [11].

The blink and non-blink regions are segregated using energy detector method and perform thresholding to the blink regions alone to preserve the neural information at non OAs zone. Table 1 delineates the performance metrics of various artifact removal methods over blink regions

Improvement in ARR specifies the extent to which artifacts are removed from the original EEG signal and the improvement in CC indicates the similarity are relatedness between the raw and clean EEG signals during blink regions. An effective artifact removal method should maintain high ARR and poor CC between the raw and clean EEG signals over the blink regions. From Table 1, it is observed that SWT is exceptionally good, whereas DWT is more applicable next to SWT.

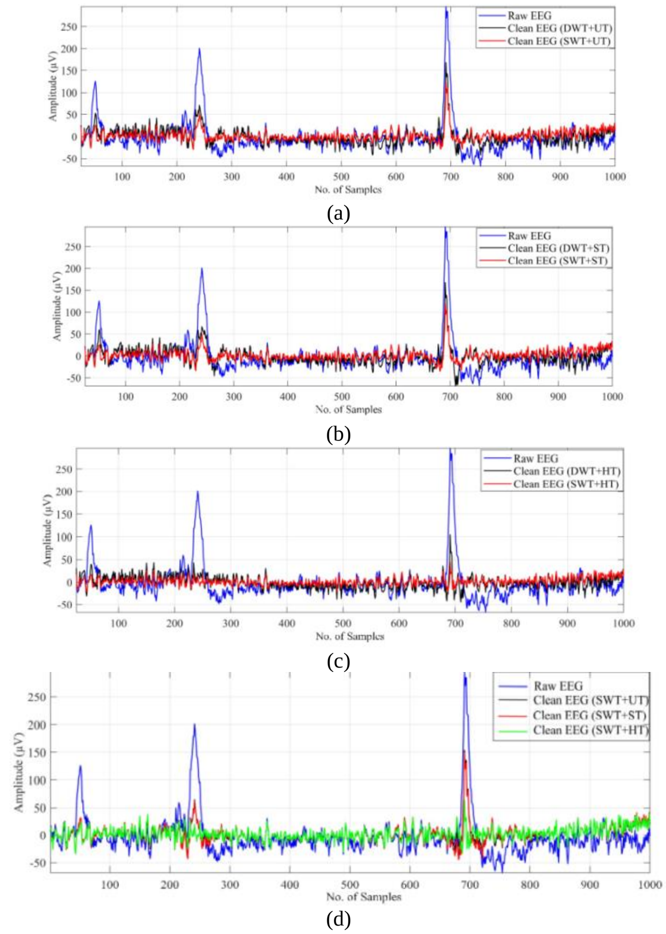


Fig. 3. (a) EEG Signal Extraction using Time Frequency Distributions (TFD), (b) EEG Signal Extraction using Fast Fourier Transform (FFT), (c) EEG Signal Extraction using Eigenvector Methods (EM), (d) EEG Signal Extraction using Stationary Wavelet Transform (SWT).

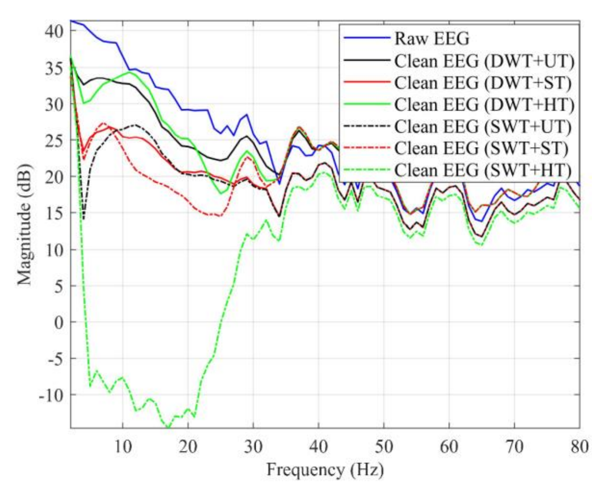


Fig. 4. Signal Analysis using Wavelet Theory DWT, UT, ST, SWT & HT.

TABLE. I. ARR AND CC BETWEEN RAW AND CLEAN EEG SIGNALS OVER BLINK REGIONS BY WT METHODS

Channel	Method	Threshold Function	Blink1 ARR	Blink1 CC	Blink2 ARR	Blink2 CC	Blink3 ARR	Blink3 CC	Blink4 ARR	Blink4 CC
F7	DWT	UT	3.22	0.422	3.62	0.362	4.17	0.245	4.94	0.236
		ST	3.84	0.384	4.26	0.286	4.64	0.228	5.16	0.198
		HT	5.98	0.312	6.14	0.243	6.4	0.212	6.84	0.175
	SWT	UT	3.92	0.254	4.54	0.238	4.84	0.193	5.18	0.198
		ST	4.48	0.214	5.22	0.205	5.24	0.171	5.56	0.156
		HT	6.26	0.152	6.98	0.118	7.22	0.142	7.6	0.122
F8	DWT	UT	2.82	0.558	3.3	0.453	3.37	0.362	4.14	0.272
		ST	3.87	0.514	3.64	0.358	3.94	0.25	4.82	0.192
		HT	5.48	0.36	6.24	0.286	5.72	0.196	6.24	0.128
	SWT	UT	3.08	0.428	3.76	0.34	3.76	0.298	4.86	0.224
		ST	4.42	0.415	4.86	0.307	4.92	0.227	5.84	0.182
		HT	6.4	0.294	7.18	0.196	7.45	0.143	8.11	0.116
Fp1	DWT	UT	2.54	0.485	3.18	0.407	3.82	0.305	4.48	0.257
		ST	3.82	0.398	4.38	0.358	4.42	0.283	5.2	0.212
		HT	6.66	0.296	6.8	0.254	7.15	0.182	7.42	0.136
	SWT	UT	3.66	0.412	4.96	0.345	5.18	0.275	5.52	0.214
		ST	4.62	0.312	5.44	0.296	5.84	0.242	6.88	0.197
		HT	5.87	0.222	6.98	0.214	7.44	0.147	7.98	0.132
Fp2	DWT	UT	2.42	0.428	2.78	0.363	3.25	0.242	4.92	0.218
		ST	3.44	0.384	4.1	0.228	4.42	0.165	5.4	0.156
		HT	6.37	0.21	7.44	0.166	7.82	0.143	8.22	0.134
	SWT	UT	4.24	0.325	4.82	0.23	5.12	0.218	5.28	0.207
		ST	4.98	0.294	5.52	0.217	5.84	0.196	6.14	0.164
		HT	6.54	0.163	6.98	0.146	7.84	0.125	8.45	0.118

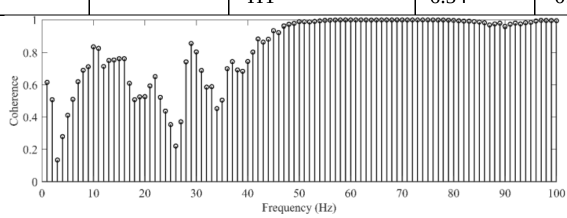


Fig. 5. Power Spectra of Raw and Clean EEG Signals by DWT and SWT Methods using Various Threshold Functions.

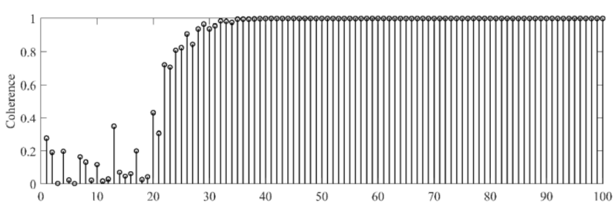


Fig. 6. Magnitude Square Coherence for Fp1 EEG Signal by a) DWT+HT b) SWT+HT.

However, threshold function HT is the best in both the methods and ST is the second best.

The average execution time required to blink region by different methods is given in presented below in Table 2 [13]. The results might be due to its larger redundancy at each decomposition level.

TABLE. II. ARR AND CC BETWEEN RAW AND CLEAN EEG SIGNALS OVER BLINK REGIONS BY WT METHODS

Method	**DWT**	**SWT**
Average Execution Time (Sec)	0.025	0.275

VII. CONCLUSIONS

In this manuscript, a Hybrid method is proposed for detection and correction of blink artifacts in single channel EEG signals. Due to finite energy difference between EEG and blink artifacts, Energy Detection method should be an optimum choice for detection of blink artifacts at various levels of EEG signal. The efficacy of the WT methods using various threshold functions are compared in terms of metrics ARR and CC during blink regions. It was observed that SWT method has shown superior performance than DWT method, and threshold function HT is better than other threshold functions in both the methods. SWT+HT method is exceptionally good in rejecting the artifacts but time consuming. Hence, DWT+HT is a better choice for correction of OAs for real time application whereas SWT+HT is the better choice for offline applications.

REFERENCES

[1] V.Krishnaveni, S.Jayaraman, S.Aravind, V.Hariharasudhan, K.Ramadoss, "Automatic Identification and Removal of Ocular Artifacts from EEG using Wavelet Transform", MEASUREMENT SCIENCE REVIEW, Volume 6, Section 2, No. 4, 2006, pp. 45-57.

[2] P. Senthil Kumar, R. Arumuganathan, K. Sivakumar, C. Vimal, "Removal of Ocular Artifacts in the EEG through Wavelet Transform without using an EOG Reference Channel", Int. J. Open Problems Compt. Math., Vol. 1, No. 3, December 2008, pp. 188-200.

[3] G.Geetha, Dr.S.N.Geethalakshm, "De-Noising Of EEG Signals Using Bayes Shrink Based on Coiflet Transform", Proc. of Int. Conf on Advances in Recent Technologies in Communication and Computing 2011, IET, pp. 41-46.

[4] Qinglin Zhao, Bin Hu, Yujun Shi, Yang Li, Philip Moore, Minghou Sun, Hong Peng, "Automatic Identification and Removal of Ocular Artifacts in EEG—Improved Adaptive Predictor Filtering for Portable Applications", IEEE Transactions on Nano-Bioscience, Vol.13, No.2, June 2014, pp. 109-117.

[5] Charvi A. Majmudar, Ruhi Mahajan, and Bashir I Morshed, "Real-Time Hybrid Ocular Artifact Detection and Removal for Single Channel EEG", 2015 IEEE International Conference on Electro/Information Technology (EIT), 21-23 May 2015, pp. 330-334.

[6] Jianbo Gao, Member, IEEE, Hussain Sultan, Jing Hu, Member, IEEE, and Wen-Wen Tung, "Denoising Nonlinear Time Series by Adaptive Filtering and Wavelet Shrinkage: A Comparison", IEEE Signal Processing Letters, Vol. 17, No. 3, March 2010, pp. 237-240.

[7] EEG Motor Movement/Imagery Dataset (eegmmidb).[Online]. Available: https://physionet.org/cgi-bin/atm/ATM, accessed 2018

[8] Harry Urkowitz, "Energy Detection of Unknown Deterministic Signals", PROCEEDINGS OF THE IEEE, APRIL 1967, pp. 523-531.

[9] Fadel F. Digham, Mohamed-Slim Alouini, MarvinK. Simon, "On the Energy Detection of Unknown Signals Over Fading Channels", IEEE Transactions on Communications, Vol. 55, No. 1, January 2007, pp. 21- 24.

[10] Hossein Shahabi, Sahar Moghimi, Hossein Zamiri-Jafarin, "EEG Eye Blink Artifact Removal by EOG Modeling and Kalman Filter", BMEI 2012, pp. 496-500.

[11] D. Hamidian, J. Salajegheh, E. Salajegheh, "Damage Detection of Irregular Plates and Regular Dams by WaveletTransform Combined Adoptive Neuro Fuzzy Inference System", Civil Engineering Journal, 2018, Vol 4, No.2, pp. 305-319, DOI: http://dx.doi.org/10.28991/cej- 030993

[12] Drapaca C.S., "Mathematical Modeling of a Brain-on-a-Chip: A Study of the Neuronal Nitric Oxide Role in Cerebral Microaneurysms", Emerging Science Journal, Vol 2, No 6 (2018), DOI: 10.28991/esj- 2018-01156

[13] Espejel-García D, Luis Ricardo Ortíz-Anchondo, Cornelio Alvarez-Herrera, Alfonso Hernandez-López, Vanessa Verónica Espejel-García, Alejandro Villalobos-Aragón., "An Alternative Vehicle Counting Tool Using the Kalman Filter within MATLAB", Civil Engineering Journal, Vol 3, No 11 (2017), DOI: 10.28991/cej-030935

[14] Garimella, Bhaskar & Shenoy, Sandeep & Pai, Rashmi & Shetty, Rishali. (2018). Structured Education for Sustainable Employment: Technology Enabled Queueing Theory Applications. Foundations of Management. 10. 75-86. 10.2478/fman-2018-0007.

Usability of Mobile Assisted Language Learning App

Kashif Ishaq*,[1], Nor Azan Mat Zin[2], Fadhilah Rosdi[3]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Malaysia

Adnan Abid[4], Qasim Ali[5]
School of Systems and Technology (SST)
University of Management and Technology
Lahore, Pakistan

***Abstract*—The aim of this study is to evaluate the usability of Mobile Assisted Language Learning i.e. Literacy and Numeracy Drive (LND) which is smartphone application to learn language and mathematics in public sector primary schools of Punjab, the biggest province of Pakistan. In this study, usability tests were conducted which included surveys of questionnaires from teachers and students. The user experience, reliability, and performance of mobile application assessed, along with user satisfaction. The LND mobile application has not been found to be successful, with a poor user interface and requires improvement. The "Using Experience," "Ease of Use" and "Usefulness" variables have been the lowest scorers in terms of user experience. Mobile device specifications were not simple and confusing; the services provided by the LND were not appealing and effective for students or teachers. This research suggested several improvements in the usability and functionality of this LND application based on assessed user experience. Many schools have chosen to use mobile apps for the teaching and evaluation of language at school. The use of mobile-assisted learning at public sector schools in Punjab, invites us to gauge the usability and effectiveness of this approach at such a huge scale which will make it more effective.**

Keywords*—Literacy and numeracy drive; usability; user experience; mobile app; assessment; public school

I. Introduction

Ubiquity and Accessibility of Mobile Phone and Internet: The growing popularity of the digital technology of the newer generation, the growth of mobile data and the fall of smartphone prices have created a strong basis for the use of smartphones on the web. This has led to a growing global number of mobile phone users. [25] announced 1375.5 million smartphone units were exported by the sellers worldwide. However, the sellers have delivered for the very first time over 1 billion units in 2013, 38.4% above last year's numbers.

In Chinese smartphone sales, the same pattern was observed. In 2013, 423 million smartphones were exported [37]; this estimate is 64.1% greater than last year, with the number of Chinese mobile users rising steadily. A total of 527 million units of Chinese mobile telephones were developed by the [9] in July 2014 and the 34th China Internet Network Development Statistical Reports were released and in 2014-15, the number of mobile subscribers in Pakistan rose to 114.7 million in 2019, about 72 million wireless subscribers and in September of 2019 more than 3G/4 G subscribers and more than 74 million subscribers with broadband services [42]. In China [8], 86 percent of university students had smartphones, and only 02% had regular mobile phones with no internet browsing capabilities.

Mobile-based learning: M-learning has various aspects, offers cost-effective learning and expands learner space [59]. This means that computing is a valuable tool in education through the development and growth of mobile technology in the various mobile telecommunications industries, which are actually the principal technology in use [14]. The key advantage of mobile learning is that students use these smart devices for learning purposes and connectivity quickly, easily and with the rapid progress of smartphone technology [45].

LND an initiative of Government of the Punjab, Pakistan: Keeping in view the benefits of M-learning, The Government of Punjab initiated a mobile-based learning application named Literacy and Numeracy Drive (LND) for the learning and assessment of public sector school students of grade 03. This project was launched in 2015, in whole province of Punjab having 36 Districts, in which 52,394 public sector schools, 403,172 teachers and 12,268,981 students enrolled [41] to eliminate the old-fashioned procedures of large scale pupil assessment such as Punjab Examination Commission (PEC) are expensive, occasional, and complex [58].

Thus, a low-cost, tablet-PC pupil evaluation application (Fig. 1) was incorporated by school monitoring personnel during their monthly visits to education by the School Education and Punjab Information and Technology Board. The evaluation app is connected to a comprehensive question bank and the related student learning outcomes are identified with each question [32]. Main student learning outcomes (KLOs) are currently evaluated for grade 3 learners for English (Fig. 2), Urdu (Fig. 3) and Maths (Fig. 4). The key areas measured were comprehension (Fig. 5), completing sentences, addition, and subtraction of two and three-digits, multiplication (Fig. 6), and division.

The students are evaluated by a monitoring and evaluation team personnel (MEA) who also uses an android-app. A tablet is used for evaluation whereby questions for each class, subject, and main SLO are randomly rendered from a central question-bank. The student to be tested are presented with these multiple-choice questions (Fig. 5). Through exam in over 47,000 public schools in Punjab brings together a total of seven questions. It takes less than five minutes per pupil to complete the process. The monthly assessment takes place at nearly 329,000 participants. Up to now, the MEAs carried out almost 6.7 million evaluations (Fig. 7). Information is exchanged via an Internet dashboard and SMS alerts [32] with training administrators.

*Corresponding Author

Fig. 1. LND Interface.

Fig. 2. Interface for English.

Fig. 3. Interface for Urdu.

Fig. 4. Interface for Math.

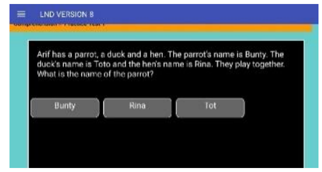


Fig. 5. LND English Test.

Fig. 6. LND Math Test.

Fig. 7. LND Test Result.

Motivation of this Research

Due to the continuous popularization and increase in the usage of smartphones, patrons prefer to use the enhanced features of mobile applications for their studies [26] [2]. Therefore, this study emphases on the LND app (Shown in Fig. 1) adopted in public sector schools of Punjab province. Since its commencement, this app has been downloaded & installed in 52,394 schools and in use for learning and assessment but till now, no study has been made for its usability either this app encounters a problem or it needs improvement. Researchers intend to assess whether or not the LND app can be stably operated on various terminals in the present study. Current usability studies have been recognized as important in the evaluation of goods and systems [21] but for the LND app, no test of usability has been developed since its enforcement. Therefore, these tests need to be conducted among educators to collect feedback and counterpoints and to recognize how students and teachers feel while visiting the LND mobile app at school. In order to measure performance, efficiency and user satisfaction, utility analysis is performed on the LND App, which is implemented by schools of Punjab public sector [60].

The rest of the article has been structured as follows. Section 2 presents the relevant literature. While the methodology used to evaluate the usability and effectiveness of LND has been presented in Section 3. The analysis and results of the study have been presented in Section 4. Discussion on the results and recommendations for improvements have been presented in Section 5. Lastly, Section 6 concludes the article.

II. Literature Review

A. Usability

Usability was initiated at the end of the eighties [7] and is commonly used in measuring product and services quality and acceptance [54], [17]. While usability study results might not be the most efficient strategy for site assessment, they provide dependable quantitative projections for user satisfaction and performance [1]. In fact, existing mobile usability studies [57], [22] have identified nine generic usability attributes: User satisfaction, Simplicity, learnability, effectiveness, errors, efficiency, memorability, comprehensibility, and performance of learning. These usability characteristics help to evaluate the quality of certain products and user-friendliness. The current usability research is among these qualities, aimed at measuring user experience, efficiency and quality with LND apps used in public sector schools of the province of Punjab.

Such three attributes are frequently taken to check the user experience of the mobile devices used for learning purpose [5], [12], [13], [15], [31]. Effectiveness means integrity and consistency in the achievement of certain objectives by specified users. Efficiency corresponds to the fast and effective manner in which users can perform the task or the tools they need to accomplish their objectives with respect to reliability and completeness [51], [53], [47]. Satisfaction with users reflects users ' position on the use of mobile applications.

B. Usability Testing

Small tests are usually used in usability research. Five participants were found to have a total of 85% of usability issues, especially when these users are [39], [20]. Nonetheless, in recent studies, large numbers were employed to study the functionality of mobile apps in light of the diversity of mobile devices and platforms. Twelve participants, for example, were requested to check the functionality of their website [40], [53], [18], 12 more were willingly hired to evaluate the usability of another related website [55], [36] and to a re-designed mobile portal of the library of Health Sciences was reclassified to 10 participants [44], [35].

In order to measure the functionality of mobile applications two main methodologies [57], namely field studies and laboratory experiments, are implemented. Field experiments are favored to laboratory tests because the actual and changing world in which operations are usually carried out is preserved. The user experience acquired in a real-world situation is, therefore, more accurate and practical than in a laboratory environment. Nonetheless, three main difficulties in performing field studies have been identified in the literature [16]. Real studies, far from trivial to use established assessment techniques, were difficult to establish and data collection was not easy. Laboratory studies often benefit from several benefits relative to field studies, including complete control over tasks, simple performance assessments and the ability to capture usage patterns using video [57].

The significance or utility of experimental and mobile research variations was not defined by agreement in previous literature [4], [43], [56]. Nonetheless [56] claimed that field research is better suited for the usability testing of mobile applications. [34] concluded that repetitive and expensive conventional laboratory testing is often incapable of representing practical uses. [53], [54] stated that lab experiments may be helpful but cannot test people's behaviors in conditions in real life. Actual mobile apps should, therefore, be used as far as possible [57]. It is important to determine the actual experience and behavior of consumers with a real environment because such interactions help to improve the mobile LND application.

C. Usability of the Mobile Educational Applications

We reside in the mobile technology era and numerous people have smartphones which are often used for internet surfing. Worldwide, five billion people use mobile phones. The number is impressive and a new mobile segment is increasing rapidly including tablets and phones. Kids use items like the iPad and tablet [24]. Their use is widespread among kids. Digital education is the way we learn, everywhere and wherever. Mobile devices support it, and it includes student and content mobility in the sense that it is available from everywhere or at any time [10]. Most software apps are commercially available for small children, and the growing popularity of digital teaching for childhood has brought about a fresh wave of mobile education. In the United States, 88 percent of public schools have an appropriate student usage plan for transferring telephones, according to a study [10], [50].

The usability of mobile apps is an important part of it and it can also be prevented through properly designed software [29]. Usability testing for mobile apps is an evolving research field, which is difficult because of its unique features, such as the smaller screen width, restricted input space, and shifting user contexts [22], [29].

A detailed review of the functionality of the mobile learning software was made in response to the disappointment voiced by students at the National University of Fiji. The study was carried out by 30 students who tested the system to make it simple to use, efficient and satisfactory. The findings showed minor problems in usability and suggestions for further development [27], [28]. The researchers analyzed students ' understanding of the effectiveness of Google Apps for Education (GAFE), how it was challenging or simple for students to communicate with GAFE and how students perceived software as a specific benefit in terms of accessibility and efficiency [6], [28].

A Primo discovery tool usability test was performed which is a medium-sized library research tool to detect patterns in user search behaviors. The researchers explored key elements of the design and features of Primo on the basis of specific research questions. In order to understand the usability of Primo for users, a diagnostics usability evaluation was conducted. Display behavior, graphical gestures and verbal commentaries were analyzed in investigators, which helped to test software usability. The study concluded many problems faced by users (respondents who misunderstood themselves in searching primo features, query boundary labels) and technologically (contradictory results compared to search indexing methods, documents obtained from digital library in-house were not shown properly) [38], [27].

CAMEG is created to make learning more appealing to the students in an interactive game-based mobile application. The researchers measured the utility of student interpretation of the play. The researchers have taken Management Information System (MIS) courses in a game developed a digital science park in 03 labs and questioned approximately 23 groups of between three and four pupils to fill out the CAMPRG usability survey. The researchers have collected user reviews and proposed that software be best used [33], [28].

MOSAD was developed using a template material from the Universiti Teknologi PETRONAS (UTP) course System Analysis, Development and Science (SDS). In order to improve efficiency and usability, a heuristic analysis with five experts in the field of human-computer interaction (HCI) was carried out, accompanied by the Post Analysis Quasi-experimental design performed by 116 UTP participants of second years testing the efficacy and usability of MOSAD in the field of SAD. Findings above 3.5 for each compound, based on the rating 1–5, were also given by the usability test which measured the uniqueness, the acquirability, the durability, the minimal behavior and the minimum memory load. The results thus show both that MOSAD is beneficial and reliable for tertiary education pupils as an instrument for the revision of education [23], [27].

As far as our study is concerned, usability is an important aspect to make the game-based learning especially game-based mobile learning effective or efficient. Hence, usability is a key quality attribute for learning. The success of game-based mobile learning requires positive student attitudes to mobile-learning, together with suitable usability and good user experience in the systems [49]. With the use of better technology in smarts phones, a number of mobile applications have developed but in most of the applications, usability is not primarily focused, which is the main reason for the failure of such applications [48].

Literacy and Numeracy Drive (LND), a well-known mobile-based application adopted by public sector schools in the Punjab province for grade 03 students (male and female). Given its widespread use in over 52,000 schools [41], though, no research was conducted on the usability of this mobile app. Usability work leads to enhancing the user experience of mobile LND apps and the approach and assessment findings offer useful guides to enhance participant and technical user experience to achieve good outcomes.

III. Methodology

A. Setting and Sample

This study took place in public sector primary schools of Pakistan which is consisted of four provinces, Punjab, Khyber Pakhtunkhwa, Balochistan, and Sindh. The Punjab province is the most populated area of Pakistan selected for this study which is further divided into 36 districts. It has 52,394 schools (male and female), 12,268,981 students (male and female) and 403,172 teachers (male and female). Furthermore, District Sheikhupura is selected for the survey and interviews of the teachers, which is sub-divided into 05 regions (Ferozwala, Muridke, Safdarabad, Sharaqpur, and Sheikhupura) and students for survey only from the same region. It has 1,247 schools (male and female) out of which 21 schools were randomly selected to visit in whole regions. This study is to measure the usability of the LND app, from the teachers and students in the school who were using LND app in the class for teaching and students who were using LND app in class and/or at home for learning. The sample was drawn from teachers of 21 randomly selected schools out of 1,247 from District Sheikhupura and 57 teachers who were teaching English, Math and Urdu subjects, also using LND application in the class were invited to participate in the study by accomplishing questionnaire and interview in class timing. A strength of 300 students who were using the LND application at home and at school was invited to complete the questionnaire. The rate of response was 100% because all teachers and students participated happily.

B. Participants

For the study of teachers, 57 respondents participated from which male respondents were 12 (21.1%) and the majority of respondents were female which was 45 (78.9%). The age group of 18 (31.6%) participants were 26-30 which is the majority whereas 09 (15.8%) respondents having the age group of 31-35 whilst 13 (22.8%) and 17 (29.8%) respondents have the age group of 36-40 and above 40 respectively.

For the study of students, 300 respondents participated from which male respondents were 138 (46.0%) and the majority of respondents were female which was 162 (54.0%). The age group of 06 (2.0%) participants were 05-07, the overwhelming majority of 217 (72.3%) participants belonged to 08-10 age group whereas 76 (25.3%) participants having the age group of 11-13 whilst only 1 (0.3%) participant was above 13 of age.

C. Instrument

The questionnaire includes a total of 40 questions was developed and reliability was measured by Cronbach's alpha which was (.856) that indicates a high level of internal consistency. The questionnaire includes five factors: demographic information, User Experience, Ease of Use, and Usefulness of application was measured. In the students' questionnaire having 30 questions that include: Demographic information, Availability of electronic devices, User Experience, Usefulness, and Ease of use for LND application was measured. The reliability of the questionnaire was measured by Cronbach's alpha which was (.805) thus indicating that the questionnaire survey was vastly reliable and stable.

For the teachers section of demographic information: age, gender, education, and location of school was asked from the participants whereas in the section of Usability and User Experience: icon of application is easy to find, icons and buttons are eye-catching and identifiable, Color scheme of button is attractive, Interface is easy to use, font size is easy to read, easy to use touch screen, easy to use input, step by step assistance, difficulty, performance and improvement, navigation keys, voice instructions, animations, videos, recommendations of questions, question bank and advertisements; in the section of Usefulness: enhance of vocabulary, saves my time, pictorial presentation, improvement of knowledge, skillful in learning and confident in speaking was asked where all items based on 5-point Likert scale, ranging from 1-Strongly disagree to 5-Strongly agree.

For the students' section of demographic information: age, gender, residential area, and location of the school was asked from the participants where the language was the first question then the availability of devices like mobile phones, smartphone, tablets, laptop, and computers at home or at school was asked. In the next section, the Availability of electronic devices, User Experience, Usefulness, and Ease of use of the LND app was asked. The questions having yes/no options were asked from the students because they were unable to make the decision on a Likert scale due to their educational capability and age group.

D. Data Collection

The data was gathered during the winter session in 2018 and the questionnaire was administered to all English, Math, and Urdu subject teachers and from the student side, only grade 03 students were focused from the public sector schools in which the participation of respondents was voluntary and anonymous. For the analysis of questionnaire data, Statistical Package for Social Science (SPSS 25.0) was used in which frequencies were generated for demographic information of the participants whereas for the sections of LND (Usability, User Experience, Usefulness, and Ease of use), Mean and Standard Deviation were calculated accordingly.

IV. Results and Discussions

A. Teachers Result

After distributing the questionnaire to the teachers in public sector schools of Punjab province, following are the results compiled by the SPSS (version 25.0) and presented in the form of a table and interpreted as well:

According to the demographic information in Table I. and Table II. majority 18 (31.6%) respondents having the age group of 26-30 whereas 09 (15.8%) respondents having the age group of 31-35. In this table, 13 (22.8%) respondents have the age group of 36-40 and 17 (29.8%) respondents have the age group of above 40. According to the gender description, out of 57 respondents, male respondents were 12 (21.1%) and the majority of respondents were female which was 45 (78.9%) of the total of 100%.

Education of the respondents is the important factor for thinking / suggesting if the respondent has higher level of education will provide appropriate answer for the question asked and the Table III. describes that 02 (3.5%) respondents have the only Matric qualification to teach students in the school whereas 03 (5.3%) respondents have Intermediate qualification and 11 (19.3%) respondents have bachelors' qualification for teaching at school level. The majority of respondents 35 (61.4%) have masters (16 years) education for teaching in schools whereas only 05 (8.8%) and 01 (1.8%) respondent has MPhil and Ph.D. (scholar) qualification for teaching. Table IV. describes that the majority of the respondents 42 (73.7%) belong to the rural area schools whereas only 15 respondents (26.3%) belong to the urban area schools.

TABLE. I. AGE

	Frequency	Percent
26-30	18	31.6
31-35	9	15.8
36-40	13	22.8
Above 40	17	29.8
Total	57	100.0

TABLE. II. GENDER

	Frequency	Percent
Male	12	21.1
Female	45	78.9
Total	57	100.0

TABLE. III. EDUCATION

	Frequency	Percent
Matric	2	3.5
Intermediate	3	5.3
Bachelor	11	19.3
Master	35	61.4
MPhil	5	8.8
Other	1	1.8
Total	57	100.0

TABLE. IV. SCHOOL AREA

	Frequency	Percent
Rural	42	73.7
Urban	15	26.3
Total	57	100.0

TABLE. V. MEAN AND STANDARD DEVIATION FOR USABILITY OF LND

Sr. #	Items	N	M	SD
1	The application icon is easy to find.	57	2.44	1.09
2	The application interface is easy to use.	57	2.25	.61
3	The application provides easy to use touch screen input.	57	1.56	.50
4	The application provides step by step assistance to use it.	57	2.14	.72
5	The application provides assistance in difficulty.	57	1.45	.50
6	The application instructs to fix the problem automatically.	57	1.15	.37
7	The application shows too many advertisements.	57	4.91	.29
8	The application provides variety of questions in its question bank.	57	1.19	.40
9	The application provide self-recommendations for questions.	57	1.02	.13
10	The application provider is taking steps to improve the application.	57	1.02	.13

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

Usability of LND: The majority of the respondents in Table V show the lowest mean score for all items except item 07. The role of usability in mobile applications considered vital because better the usability leads to the smooth and effective use of an application. From the item 1 to 10 in the table depicts the problems faced by the respondents while using LND application in which item 09, and 10 has the lowest mean score for non-availability of self-recommendations questions and steps to improve the application for ease to use by the respondents. Item 03 depicts the problem of input by (90.3%) majority where ($\overline{X} = 1.56\ and\ SD = .50$) and no assistance availability by (93.5%) majority where ($\overline{X} = 1.45\ and\ SD = .50$) for the difficulty faced by respondents in item 05. Advertisements in an application have a bad impression and it affects the usage of the application for the meaningful purpose which is described in item 7 (highest mean score), where (91.2%) respondents ($\overline{X} = 4.91\ and\ SD = .29$) strongly agrees that the LND application displays so many advertisements that distract the process of teaching and learning of the respondents.

Ease of Use: From items 11 to 20 of Table VI shows that respondents have the problem for using the LND application easily and efficiently. Font size is a major element for consideration while reading text on the mobile-based application and if the font is not appropriate for reading then it is difficult to use such application. For the purpose, item 12 in the table, explains that the majority of respondents (80.7%) where ($\overline{X} = 1.94\ and\ SD = .44$) are not agreed with the font size used in the LND application because the respondents felt difficulty while reading the text from the screen of the application. The performance of application also has an effect on the smooth and motivational use of it because if the application performance is slow while performing the task then the user will not learn effectively in due time. Item 17, in which (98.4%) majority where ($\overline{X} = 4.75\ and\ SD = .43$) strongly agreed that LND application performance is slow while using it into the class.

For items 18 to 20, the mean score is the lowest for the voice instruction, animations and videos for learning because these elements motivate the respondents for better and ease of learning with better results. In these items the overwhelming majority (100%) where ($\overline{X} = 1.00\ and\ SD = .00$) strongly disagreed that the LND application provides voice instructions, animations and videos for learning.

TABLE. VI. MEAN AND STANDARD DEVIATION FOR EASE TO USE OF LND

Sr. #	Items	N	M	SD
11	The application is difficult to use.	57	3.78	.54
12	The font size is easy to read.	57	1.94	.44
13	The application provides navigation keys.	57	1.14	.35
14	The icons and buttons are attractive and recognizable.	57	1.70	.57
15	The color scheme of buttons is attractive.	57	1.95	.23
16	The color scheme of application screen is attractive.	57	1.56	.50
17	The performance of application is slow.	57	4.75	.43
18	The application provides useful voice instructions.	57	1.00	.00
19	The application provides animations for learning.	57	1.00	.00
20	The application provides videos for learning.	57	1.00	.00

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

Usefulness of LND: The purpose of the mobile-based application is to enhance the knowledge for meaningful purpose with specific content. But according to Table VII for LND usefulness, all the items have low mean score and item 01 describes that (96.5%) majority respondents where ($\overline{X} = 1.54\ and\ SD = .75$) disagreed that the LND application enhances the vocabulary of the respondents. The pictorial representation of concept makes the topic easy and more useful that motivates the respondent towards its use but in this table, the lowest mean score is with item 03 in which the majority (93.0%) respondents where ($\overline{X} = 1.07\ and\ SD = .25$) disagreed that the LND application provides any pictorial presentation of concepts.

Equality is necessary for using technology between students in the class, so every student can get proper time with equal access to the application to learn. Item 07 and 08 describes that (96.5%) majority where ($\overline{X} = 1.61\ and\ SD = .82$) disagree that each student gets equal access & time for the practice of the LND application in the class because every school is allowed to have only 01 tablet in the huge strength of the class. The weakness is an element that helps a respondent to improve it after getting some evaluation using some technology automatically. If the application provides such feature then it will be much effective towards interactive and efficient usage but in LND, the overwhelming majority of the respondents (98.5%) where ($\overline{X} = 1.25\ and\ SD = .43$) disagreed that the LND application provides weakness of the students on the basis of results.

B. Students Result

After distributing the questionnaire to the students in public sector schools of Punjab province, in which some teacher helped students to solve the questionnaire according to their answers, following are the results compiled by the SPSS (version 25.0) and presented in the form of the table and interpreted as well.

According to Table VIII. Age of only 06 (02%) respondents was in the group of 05-07 whereas the majority of the respondents 217 (72.3%) were in the age group of 08-10. In the age group of 11-13, 76 respondents which are (25.3%) and lastly only 01 respondent (0.3%) belongs to the above 13 age group. According to Table IX. Out of 300 respondents, male students were 138 (46%) and the majority of respondents were female which were 162 that is 54% for a total of 100%.

According to the Table X. only 18 (6%) respondents have the tablet at their home whereas majority of the respondents 227 (75.9%) have simple mobile phone at their homes to use but 186 (62.2%) majority of respondents have smart phones at their homes to use for playing games or educational purpose. Only 16 respondents (5.4%) and 100 (33.4%) respondents have laptop and computers respectively at their homes to use it for playing games, watching cartoons or educational activities.

According to Table XI, in the schools, 100% of respondents have tablets and computers to use in educational activities because Government of Punjab provides 01 tablet and 16 computers to every public school in the whole province whereas the access of computer is only for Grade 9th and 10th

students for their learning activities according to the instructions of teachers.

According to Table XII. In which only 72 respondents have computers or access to computers at their homes further responded by 67 (93.1%) respondents who use a computer for playing games at home and 58 (80.6%) respondents use a computer for watching cartoons but only 8 (11.1%) respondents use computer for their homework at homes.

TABLE. VII. MEAN AND STANDARD DEVIATION FOR USEFULNESS OF LND

Sr. #	Items	N	M	SD
21	It helps me to enhance my vocabulary.	57	1.54	.75
22	It saves my time when I use it for teaching.	57	1.47	.50
23	The application provides pictorial presentation of concepts.	57	1.07	.25
24	The application improves my knowledge.	57	1.75	.51
25	The application makes me skillful in learning English.	57	1.84	.65
26	The use of application makes me confident in speaking English.	57	1.21	.41
27	Each student gets equal access to the application in class.	57	1.61	.82
28	Each student gets equal time for practice of the application in class.	57	1.63	.70
29	Results in application helps in improving learning outcomes.	57	1.28	.49
30	The application provides the weakness of students.	57	1.25	.43

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

TABLE. VIII. AGE OF RESPONDENT

	Frequency	Percent
05-07	6	2.0
08-10	217	72.3
11-13	76	25.3
Above 13	1	.3
Total	300	100.0

TABLE. IX. GENDER

	Frequency	Percent
Male	138	46.0
Female	162	54.0
Total	300	100.0

TABLE. X. ICT_HOME FREQUENCIES

		Responses	Percent of Cases
		N	
1	Do you have Tablet at home?	18	6.0%
2	Do you have Mobile Phone at home?	227	75.9%
3	Do you have Smart Phone at home?	186	62.2%
4	Do you have Laptop at home?	16	5.4%
5	Do you have Computer at home?	100	33.4%

TABLE. XI. ICT_SCHOOL FREQUENCIES

		Responses	Percent of Cases
		N	
6	Do you have Tablet at school?	300	100.0%
7	Do you have Computer at school?	300	100.0%

TABLE. XII. COMPUTER_USE FREQUENCIES

		Responses	Percent of Cases
		N (72)	
8	Do you use computer for Games?	67	93.1%
9	Do you use computer for Homework?	8	11.1%
10	Do you use computer for Cartoons?	58	80.6%
11	Do you use computer for Facebook?	1	1.4%
12	Do you use computer for Music?	1	1.4%

In Table XIII, only 139 (46.3%) respondents responded that they have access to smartphones at home to use for playing games or watching cartoons whereas 161 (53.7%) respondents do not have access to smartphones at their homes. to know the frequency and percentage of using smartphones for playing games only, 138 (46.0%) respondents use the smartphone for playing games but 162 (54.0%) respondents do not use smartphones for playing games.

According to Table XIV, the item 15 describes that only 62 (20.7%) respondents think LND software is suitable for learning English but majority of the respondents 238 (79.3%) do not think LND is suitable for learning English because overwhelming majority 283 (94.3%) of respondents think the LND application is not interesting. The interactive screen with suitable color scheme attracts the users to use the mobile application for learning but respondents said that LND application does not have an interactive screen, colors scheme, navigations and animations whilst 280 (93.3%) of respondents responded that they cannot learn English easily with the help of LND application. The curriculum of mobile-based application must be suitable and appropriate for learning topics effectively but it is reported by 283 (94.3%) respondents that the content used in the LND application is not suitable.

The suitable content/curriculum defined by Government department/officials is necessary for the learning of topics in the class but in the absence of suitable content, there will be a bad effect on learning as well as the assessment of pupils. The item 20 and 21 describes that majority 289 (96.3%) of respondents cannot learn comprehension easily with the help of LND application because this application is not providing suitable content automatically which is also proved by 100.0% of majority. This is also the reason that 99.0% of the majority did not take interest to learn comprehension using this LND application.

Assessment is the necessary important element in the class to see the progress of the student and with an appropriate assessment method, the teacher may know that to what extent students learned the topic effectively. The majority of the respondents 200 (66.7%) are not agreed with the assessment method used by LND because the assessment is not taking part to improve comprehension for the English language which is responded by 283 (94.3%) respondents. The application also does not provide the content to learn comprehension on the basis of result which is justified by 100.0% of majority and 99.7% of respondents say that the LND application does not provide the weakness to learn comprehension for the pupil.

TABLE. XIII. SMARTPHONE AND USAGE

Sr. #	Item	Frequency		Percent	
		Yes	No	Yes	No
13	Do you have an access to a Smartphone?	139	161	46.3	53.7
14	Do you use smartphone for playing games only?	138	162	46.0	54.0

TABLE. XIV. USABILITY, EASE OF USE OF LND APPLICATION

Sr. #	Items	Frequency		Percent	
		Yes	No	Yes	No
15	Do you think Literacy, Numeracy Drive (LND) software is suitable for you to learn English?	62	238	20.7	79.3
16	Do you think that LND application is interesting in learning English?	17	283	5.7	94.3
17	Do you think you can learn English easily from LND application?	20	280	6.7	93.3
18	Do you think that content of English learning is suitable in LND application?	17	283	5.7	94.3
19	Do you think teacher teaches English well with the help of LND application?	156	144	52.0	48.0
20	Do you think you can learn comprehension easily with the help of LND application?	11	289	3.7	96.3
21	Do you think application provides you content to learn comprehension automatically?	0	300	0.0	100.0
22	Do you think application develops interest to learn comprehension?	3	297	1.0	99.0
23	Do you think the method of assessment adopted in LND is suitable for you?	100	200	33.3	66.7
24	Assessment of LND improve your English learning?	17	283	5.7	94.3
25	Application provides the content to learn comprehension on the basis of results?	0	300	0.0	100.0
26	Application provides you your weaknesses to learn comprehension?	1	299	0.3	99.7

According to Table XV, the majority 283 (94.3%) respondents said that the LND application does not have interactivity in screen or buttons to attract the pupils to use this application with the passion to learn English. It is also mentioned by the majority of respondents 218 (72.7%) that content and exercises are not properly available for the practice of learning English. Animated videos and voice pronunciations attract the pupils to learn effectively and also can produce good results but according to 298 (99.3%) and 291 (97.0%) respondents, these features are missing in LND.

TABLE. XV. FEATURES TO BE IMPROVED IN LND

		Responses	Percent of Cases
		N	
27	Do you think Interactive Screen and Icons are not available in LND application for you to learn English?	283	94.3%
28	Do you think Content and Exercises are not available in LND application for you to learn English?	218	72.7%
29	Do you think Animated Learning Cartoons are not available in LND application for you to learn English?	298	99.3%
30	Do you think Voice Pronunciation is not available in LND application for you to learn English?	291	97.0%

V. DISCUSSIONS AND RECOMMENDATIONS

After conducting a survey from teachers and students to see the usability of LND in public sector schools for teaching and learning for students of grade 03, several problems were identified and the teacher also recommended some improvement for mobile-based application LND to overcome all the issues.

- Game-based feature should be introduced to enhance the user experience of students and teachers in current LND app which is also described by [19] and using game-based features, challenges also be included for the pupils to enhance the usability and to keep focus while on good and interactive design for learning application which is also mentioned in [46].

- The design of the interface should be accurate that have a good impact on the success of the learning application and to develop the reading and to encourage the reading by playing video games. This is also described by [30]. The presentation of learning material through games and animations inspire the pupils to improve comprehension and also to gain cognitive skills. This has been described by [11] in their study.

- Ease of use, usefulness, and clarity of goal in the game-based application that inspired the learners for using games. The designers must keep in mind these factors while designing and teachers also need to integrate these factors to get the results better for encouraging behavior from learners by participating in such learning activities. This also mentioned by [52] in their study.

- The feature of music in game-based learning applications should be incorporated, which helps the user to develop their motivation towards better learning. It is also necessary to add a feedback feature to improve the learning capability through a tactile observation of substances. These points are also described by [3] in their study.

VI. Conclusion

In the coming years, the development of a smartphone LND app is projected to increase exponentially. Smartphones are becoming more and more popular, and both school and house learners practice using this software by downloading them from smartphones of their parents. Usability assessments evaluate not only the performance, efficacy and customer engagement of mobile application but also provide multiple enhancement recommendations based on user experience. These results will surely make a significant contribution to designing and improving such systems. In fact, improving the functionality of the mobile application tested will also encourage people to use the LND mobile app and significantly enhance the quality of this product. There is a strong need to improve such games to increase their support in teaching and learning. Furthermore, they can lead to developing applications for personalized learning support at a massive level.

Acknowledgement

This study is funded by the Universiti Kebangsaan Malaysia (The National University of Malaysia), Bangi, Selangor, Malaysia.

References

[1] Alshamari, M., & Mayhew, P. (2009). Technical review: Current issues of usability testing. IETE technical review, 26(6), 402-406.

[2] Arif, I., Aslam, W., & Ali, M. (2016). Students' dependence on smartphones and its effect on purchasing behavior. South Asian Journal of Global Business Research, 5(2), 285-302.

[3] Arnab, S., Petridis, P., Dunwell, I., & de Freitas, S. (2011). Enhancing learning in distributed virtual worlds through touch: a browser-based architecture for haptic interaction. In Serious Games and Edutainment Applications (pp. 149-167). Springer, London.

[4] Billi, M., Burzagli, L., Catarci, T., Santucci, G., Bertini, E., Gabbanini, F. and Palchetti, E. (2010), "A unified methodology for the evaluation of accessibility and usability of mobile applications", Universal Access in the Information Society, Vol. 9 No. 4, pp. 337-356.

[5] Botha, A., Van Greunen, D. A. R. E. L. L. E., & Herselman, M. A. R. L. I. E. N. (2010). Mobile human-computer interaction perspective on mobile learning.

[6] Brown, M.E. and Hocutt, D.L. (2015) 'Learning to use, useful for learning: a usability study of Google apps for education', Journal of Usability Studies, Vol. 10, No. 4, pp.160–118.

[7] Butler, K.A. (1996), "Usability engineering turns 10", Interactions, Vol. 3 No. 1, pp. 59-75.

[8] Chaputula, A. H. (2016). eReadiness of public university libraries in Malawi with special reference to the use of mobile phones in the provision of library and information services (Doctoral dissertation).

[9] China Internet Network Information Center (2015), "Statistical report of the thirty-fourth China Internet development", available at: www.cnnic.cn/gywm/xwzx/rdxw/2014/201407/ W020140721 559080702009 (accessed November 13, 2019).

[10] Chou, C. C., Block, L., & Jesness, R. (2012). A case study of mobile learning pilot project in K-12 schools. Journal of Educational Technology Development and Exchange (JETDE), 5(2), 3.

[11] Conolly, T., Stansfield, M., Boyle, L. (2009): Games-Based Learning Advancements for Multy-Sensory Human Computer Interfaces: Techniques and Effective Practices. IGI Global Publishing.

[12] Coursaris, C.K. and Kim, D.J. (2006), "A qualitative review of empirical mobile usability studies", Proceedings of the Twelfth Americas Conference on Information Systems, Vol. 5, pp. 2873-2879.

[13] Coursaris, C. K., & Kim, D. J. (2011). A meta-analytical review of empirical mobile usability studies. Journal of usability studies, 6(3), 117-171.

[14] Davis, D., Chen, G., Hauff, C., & Houben, G. J. (2018). Activating learning at scale: A review of innovations in online learning strategies. Computers and Education, 125(August 2017), 327–344. https://doi.org/10.1016/j.compedu.2018.05.019.

[15] De Villiers, M. R., & Harpur, P. A. (2013, October). Design-based research-the educational technology variant of design research: illustrated by the design of an m-learning environment. In proceedings of the South African institute for computer scientists and information technologists conference (pp. 252-261). ACM.

[16] Duh, H. B. L., Tan, G. C., & Chen, V. H. H. (2006, September). Usability evaluation for mobile device: a comparison of laboratory and field tests. In Proceedings of the 8th conference on Human-computer interaction with mobile devices and services (pp. 181-186). ACM.

[17] Dumas, J. S., & Fox, J. E. (2009). Usability testing: Current practice and future directions. In Human-Computer Interaction (pp. 247-268). CRC Press.

[18] sFung, R. H. Y., Chiu, D. K., Ko, E. H., Ho, K. K., & Lo, P. (2016). Heuristic usability evaluation of university of hong kong libraries' mobile website. The Journal of Academic Librarianship, 42(5), 581-594.

[19] Gabriel, P., Hirashima, T., & Yusuke, H. (2018). A Serious Game for Improving Inferencing in the Presence of Foreign Language Unknown Words. International Journal of Advanced Computer Science and Applications, 9(2). https://doi.org/10.14569/IJACSA.2018.090202.

[20] Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. Field methods, 18(1), 59-82.

[21] Griggs, K., Bridges, L.M. and Rempel, H.G. (2009), "Library/mobile: tips on designing and developing mobile web sites", Code4Lib Journal, No. 8, available at: http://journal.code4lib.org/articles/2055 (accessed November 14, 2019).

[22] Harrison, R., Flood, D., & Duce, D. (2013). Usability of mobile applications: literature review and rationale for a new usability model. Journal of Interaction Science, 1(1), 1.

[23] Hashim, A.S., Wan Ahmad, W.F. and Ahmad, R. (2011) 'Usability and effectiveness of mobile learning course content application as a revision tool', Computer Technology and Application, Vol. 2, No. 2, pp.149–158.

[24] Ibrahim, N., & Ahmad, W. F. W. (2016, August). Practitioners' validation on effectiveness of multimedia Mobile Learning Application for children. In 2016 3rd International Conference on Computer and Information Sciences (ICCOINS) (pp. 103-108). IEEE.

[25] IDC (2019, May 30). IDC Forecasts Worldwide Smartphone Market Will Face Another Challenging Year in 2019 with a Return to Growth on the Horizon. Retrieved November 13, 2019, from https://www.idc.com/getdoc.jsp?containerId=prUS45115119.

[26] Jin, B. S., Yoon, S. H., & Ji, Y. G. (2013). Development of a Continuous Usage Model for the Adoption and Continuous Usage of a Smartphone. International Journal of Human-Computer Interaction, 29(9), 563–581. https://doi.org/10.1080/10447318.2012.729997.

[27] Kumar, B. A., & Mohite, P. (2016). Usability guideline for mobile learning apps: An empirical study. International Journal of Mobile Learning and Organisation, 10(4), 223–237. https://doi.org/10.1504/IJMLO.2016.079499.

[28] Kumar, B.A. and Mohite, P. (2016) 'Usability study of mobile learning application in higher education context: an example from Fiji National University', Mobile Learning in Higher Education in the Asia Pacific: Harnessing Trends and Challenging Orthodoxies, Springer, New York.

[29] Kumar, B. A., & Mohite, P. (2018). Usability of mobile learning applications: a systematic literature review. Journal of Computers in Education, 5(1), 1–17. https://doi.org/10.1007/s40692-017-0093-6.

[30] Laamarti, F., Eid, M., & El Saddik, A. (2014). An Overview of Serious Games. International Journal of Computer Games Technology, 2014, 1–15. https://doi.org/10.1155/2014/358152.

[31] Lee, D., Moon, J., Kim, Y. J., & Mun, Y. Y. (2015). Antecedents and consequences of mobile phone usability: Linking simplicity and interactivity to satisfaction, trust, and brand loyalty. Information & Management, 52(3), 295-304.

[32] LND. (n.d.). Retrieved November 13, 2019, from https://open.punjab.gov.pk/lnd/reports/aboutus.

[33] Lu, C., Chang, M., Huang, E. and Chen, C.W. (2011) 'Architecture and collaborations among agents in mobile educational game', Pervasive Computing and Communications Workshops (PERCOM Workshops), 21–25 March, IEEE, Seattle, WA, pp.556–560.

[34] Ma, X.X., Yan, B., Chen, G.L., Zhang, C.H., Huang, K., Drury, J. and Wang, L.Z. (2013), "Design and implementation of a toolkit for usability testing of mobile apps", Mobile Networks and Applications, Vol. 18 No. 1, pp. 81-97.

[35] Mi, M., Wu, W., Qiu, M., Zhang, Y., Wu, L., & Li, J. (2016). Use of mobile devices to access resources among health professions students: a systematic review. Medical reference services quarterly, 35(1), 64-82.

[36] Miller, R. E., Vogh, B. S., & Jennings, E. J. (2013). Library in an app: Testing the usability of Boopsie as a mobile library application. Journal of Web Librarianship, 7(2), 142-153.

[37] Ministry of Industry and Information Technology (2014), "Review and prospect of the development of mobile phone industry in 2013", available at: www.miit.gov.cn/n11293472/ n11293832/ 11294132/n12858462/15915247.html (accessed November 13, 2019).

[38] Nichols, A., Billey, A., Spitzform, P., Stokes, A. and Tran, C. (2014) 'Kicking the tires: a usability study of the primo discovery tool', Journal of Web Librarianship, Vol. 8, No. 2, pp.172–195.

[39] Nielsen, J. and Landauer, T.K. (1993), "A mathematical model of the finding of usability problems", Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems, ACM, pp. 206-213.

[40] Pendell, K.D. and Bowman, M.S. (2012), "Usability study of a library's mobile website: an example from Portland state university", Information Technology and Libraries, Vol. 31 No. 2, pp. 45-62.

[41] PMIU. (n.d.). Retrieved November 13, 2019, from https://open.punjab.gov.pk/schools/home/landing.

[42] PTA Subscribers. (2019, September). Retrieved November 13, 2019, from https://www.pta.gov.pk/en/telecom-indicators/1.

[43] Queirós, A., Silva, A., Alvarelhão, J., Rocha, N. P., & Teixeira, A. (2015). Usability, accessibility and ambient-assisted living: a systematic literature review. Universal Access in the Information Society, 14(1), 57-66.

[44] Rosario, J.A., Ascher, M.T. and Cunningham, D.J. (2012), "A study in usability: redesigning a health sciences library's mobile site", Medical Reference Services Quarterly, Vol. 31 No. 1, pp. 1-13.

[45] Rozario, R., Ortlieb, E., & Rennie, J. (2016). Interactivity and mobile technologies: An activity theory perspective. In Mobile learning design (pp. 63-82). Springer, Singapore.

[46] S. Deterding, (2015):The lens of intrinsic skill atoms: A method for gameful design, Human–Computer Interaction, vol. 30, no. 3-4, pp. 294–335.

[47] Sarrab, M., Elbasir, M., & Alnaeli, S. (2016). Towards a quality model of technical aspects for mobile learning services: An empirical investigation. Computers in Human Behavior, 55, 100–112. https://doi.org/10.1016/j.chb.2015.09.003.

[48] Shafiq, S., & Khan, T. A. (2018). Role & value of usability in educational learning via game based apps. International Journal of Scientific and Technology Research, 7(11), 70–77.

[49] Tahir, R., & Arif, F. (2014). Framework for evaluating the usability of mobile educational applications for children. The International Journal of E-Learning and Educational Technologies in the Digital Media (IJEETDM), 1(1), 16-31.

[50] Tahir, R. (2015). a Measurement Model Based on Usability Metrics for Mobile Learning User Interface for Children. The International Journal of E-Learning and Educational Technologies in the Digital Media, 1(1), 16–31. https://doi.org/10.17781 /p001503.

[51] Wallace, S., Reid, A., Clinciu, D. and Kang, J.S. (2013), "Culture and the importance of usability attributes", Information Technology & People, Vol. 26 No. 1, pp. 77-93.

[52] Wang, Y., Rajan, P., Sankar, C. S., & Raju, P. K. (2017). Let them play: the impact of mechanics and dynamics of a serious game on student perceptions of learning engagement. IEEE Transactions on Learning Technologies, 10(4), 514-525.

[53] Wei, Q., Chang, Z., & Cheng, Q. (2015). Usability study of the mobile library App: an example from Chongqing University. Library Hi Tech, 33(3), 340-355.

[54] Wisniewski, J. (2011), "Mobile usability", Bulletin of the American Society for Information Science and Technology (Online), Vol. 38 No. 1, pp. 30-32.

[55] Yeh, S.T. and Fontenelle, C. (2012), "Usability study of a mobile website: the health sciences library, university of colorado anschutz medical campus, experience", Journal of the Medical Library Association, Vol. 100 No. 1, pp. 64-68.

[56] Zein, S., Salleh, N., & Grundy, J. (2016). A systematic mapping study of mobile application testing techniques. Journal of Systems and Software, 117, 334-356.

[57] Zhang, D. and Adipat, B. (2005), "Challenges, methodologies, and issues in the usability testing of mobile applications", International Journal of Human-Computer Interaction, Vol. 18 No. 3, pp. 293-308.

[58] Ishaq, K., Zin, N. A. M., Rosdi, F., Abid, A., & Farooq, U. (2019), Effectiveness of Literacy & Numeracy Drive (LND): A Students' Perspective.

[59] Hashim, H., Yunus, M. M., Embi, M. A., & Ozir, N. A. M. (2017). Mobile-assisted language learning (MALL) for ESL learners: A review of affordances and constraints. Sains Humanika, 9(1-5).

[60] Ishaq, K., Zin, N. A. M., Abid, A., Rosdi, F., & Ali, Q. (2020). Usefulness of Mobile Assisted Language Learning App. International Journal of Engineering and Advanced Technology (IJEAT), 9(3).

Classification of Non-Discriminant ERD/ERS Comprising Motor Imagery Electroencephalography Signals

With Novel REP-based Approach

Zaib unnisa Asif[1], M. Sultan Zia[2]
Aneela Abbas[4,] Sadaf Ilyas[5]
Department of Computer Science and IT, University of Lahore, Gujrat, Pakistan

Umair Muneer Butt[3]
Department of Computer Science and IT
University of Lahore, Pakistan
Universiti of Sains, Malaysia

***Abstract*—Classification of Motor Imagery (MI) Electroencephalography (EEG) signals has always been an important aspect of Brain Computer Interface (BCI) systems. Event Related Desynchronization (ERD)/ Event Related Synchronization (ERS) plays a significant role in finding discriminant features of MI EEG signals. ERD/ERS is one type and Evoked Potential (EP) is another type of brain response. This study focuses upon the classification of MI EEG signals by Removing Evoked Potential (REP) from non-discriminant MI EEG data in filter band selection, called REP. This optimization is done to enhance the classification performance. A comprehensive comparison of several pipelines is presented by using famous feature extraction methods, namely Common Spatial Pattern (CSP), XDawn. The effectiveness of REP is demonstrated on the PhysioNet dataset which is an online data resource. Comparison is done between the performance of pipelines including proposed one (Common Spatial Pattern (CSP) and Gaussian Process Classifier (GPC)) as well as before and after applying REP. It is observed that the REP approach has improved the classification accuracy of all the subjects used as well as all the pipelines, including state of the art algorithms, up to 20%.**

Keywords*—*MI EEG Signals; non-discriminant ERD/ERS; evoked potentials; common spatial pattern; Gaussian process classifier

I. Introduction

Advancement in technology leads to facilitate handicapped personals in daily activities like normal persons. This can be achieved by bridging the space between machines and humans, the latest research is oriented towards using brain waves for directly interacting with computers in the form of Brain Computer Interface (BCI), without using any motor activity [1]. Brain waves are non-stationary signals generated by the brain. These are acquired by many methods; one of them is Electroencephalography (EEG). This is the first non-invasive (non-surgical) method and is very popular in recent researches [2]. For simple task detection, stimuli are observed in EEG signals by finding a change in the electrical activity of brain signals. If these signals are generated as a result of imagining any motor activity such waves are called Motor Imagery (MI) EEG signals.

This paper focuses on the study of MI EEG signals. When a person thinks about an activity, an event is generated. This event can be blinking of an eye or movement of hands, tongue or feet. Events related to motor activities are called motor movements. In the brain, these signals are produced inside the motor cortex region. Electrodes can be placed on the motor cortex region to get these signals. EEG signals can have different frequencies and depending upon these frequencies, signals are divided into various categories i.e. Delta, Theta, Alpha, Beta and Gamma. Alpha/mu and Beta are used to identify motor activity in the brain. When the frequency of alpha waves is high, the brain is in a relaxed state. When someone thinks of performing any motor activity the frequency of alpha waves decreases. This decrease in frequency is called Desynchronization. As this desynchronization occurs as a result of an event therefore it is called Event Related Desynchronization (ERD). The frequency of beta waves increases by imagining any motor activity, this is called Event Related Synchronization (ERS) [3]. These values of ERS and ERD play a vital role in the classification of EEG signals. If there are no discriminant values in ERD maps then the classification becomes a troublesome task [3]. Evoked Potential (EP) is another type of brain response, which occurs as a result of internal or any external event. EP is phase-in locked to the event and ERD/ERS is not phase-in locked to the event and it's the major difference between the both [4]. EP has become the primary tool for neuroscientists when averaging of EEG signals is practiced [2]. ERDs become primary tools while dealing with MI EEG signals.

II. Brain Responses

There are two types of brain responses, ERD/ERS and Evoked Potential (EP). ERD/ERS are frequency band-specific and not phase-in locked to the event. Subjects that have non-discriminant ERD/ERS values are almost impossible to classify [3].

EP is another type of brain response. It is a time domain signal and phase-in locked to the event. There are two types of evoked potentials, Visual Evoked Potential (VEP) that occurs as a result of any visual stimuli. Event Related Potential

(ERP), that occurs as a result of any event. This event can be a reflex action, memory updating task, cognitive processing or motor imagery [5]. Steady State Visually Evoked Potential (SSVEP) is a good example of VEP. Motor imagery is an example of Spontaneous BCI [3].

This study focuses on data from which evoked potentials are removed. A new pipeline is proposed for classifying data after removing evoked potentials. The proposed pipeline comprises of Common Spatial Pattern (CSP) and Gaussian Process Classifier (GPC). For this particular type of data (non-discriminant ERD/ERS comprising MI EEG signals), our proposed pipeline outperforms state of the art algorithms. According to our best knowledge, the proposed pipeline is not previously used in this kind of research. Discussed results are the outcome of the experiments conducted in this study and the observations that are found as a result of the regeneration of other researchers' studies. Because of the non-availability of codes, generated results may differ from other researchers. By confining signals and narrowing down the region of interest, the performance of classifiers can be dramatically improved. A comprehensive comparison of several pipelines is presented to show the robustness of the proposed approach.

In the coming section background study and literature review is presented. The proposed pipeline and approach are discussed in the next section. Experiments and Results are presented in the following section and paper is concluded by the last section.

III. Literature Review

BCI's non-usage in the real-time application is caused by poor accuracy, noise and non-stationary nature of EEG signals [3]. From the literature, it is known that for BCI based on MI EEG signals, highly depend on ERD/ERS i.e. Event Related Desynchronization and Event Related Synchronization. This occurs in alpha and beta bands of EEG signals due to motor imagery activity. Large number of studies confirms that ERD/ERS is found in mu/alpha rhyme and beta rhyme in the motor cortex region of the brain. Alpha ranges from 7 to 12 Hz and beta ranges from 13 to 30 Hz [6]. Researchers have achieved very good accuracy by using CSP as a feature extraction technique [7]. Change in movement sequence will increase mu/alpha ERD. Once new movement is learned ERD will be reduced. With the help of ERD, one can find when a person is trying to learn a new motor task. After learning, ERD will be reduced. ERS means, increase in the spectral peak for a specific frequency. An increase in the alpha band or ERS in mu band means there is no activity [4]. Visual inspection of ERD/ERS is performed to find the strongest or discriminant values [8]. Subjects in a dataset that don't comprise discriminant ERD/ERS values are considered as bad [9]. Riemannian approaches seem to meet the challenges of MI EEG classification as per the latest research. In this approach, covariance matrices are directly computed from EEG signals with the help of Riemannian geometry. It requires less data for classification which makes it robust for real-time applications. Mapping the EEG signal into a Riemannian manifold, target and non-target ERP responses could be classified directly by a non-Euclidean distance measurement. Because this method does not train a model in a high dimensional space, calibration times are significantly reduced. ERP waveform with Riemannian geometry is used with P300 signals for generating spellings of words. The author claims to reduce spelling time to be within 20 seconds [10]. On digging ancient grounds, it is found that ERP has a rich history. It is the most studied activity in EEG signals. ERP is a time-domain waveform whereas ERD/ERS is usually confined to a specific frequency. Mostly ERP is used for P300 signals and cognitive processing, to find attention. To make a robust BCI system it is important to deal with noise in EEG signals. Some authors argue that linear denoising techniques affect the salient features of EEG signals. They proposed nonlinear denoising techniques [11]. For feature extraction, it is important to capture vital signal characteristics, so it can be used to classify task-specific brain states [3]. In some BCI systems, only the most dominant features are used for classification by the classifiers [12]. Based on the classification results, the type of task is identified. Famous classification approaches used in literature are Linear Discriminant Analysis (LDA) [12] [13], Support Vector Machines (SVM) [12], k-nearest neighbors, Logistic Regression (LR) [14], Quadratic Classifiers [15], Recurrent Neural Network (RNN) [16]. Some other BCI uses feature extraction, selection, and classification as one block, in deep learning [6]. CSP + LDA, CSP + LR and CSP + SVM have been state of art algorithms for classifying motor imagery EEG signals so far [1]. According to some researchers, ERP Covariance with Riemannian geometry outperformed state-of-art algorithms [17]. Fig. 1 shows the usage of MI EEG data in the literature. Most of the times EEG raw data is used for MI EEG signals as per the literature review given in Table I. Classification pipeline (Covariance matrices + Riemannian Geometry) is followed by winners of many international competitions for several years [17] therefore gained popularity in research [18] [19]. A Riemannian Geometry algorithm Minimum Distance to Mean (MDM) is famous in classifying P300 signals as well [20]. Multiple Time Window LDA Beam former (MTWLB) was used as feature extraction technique which extracted features from ERP signals and gave accuracy of 92% [21].

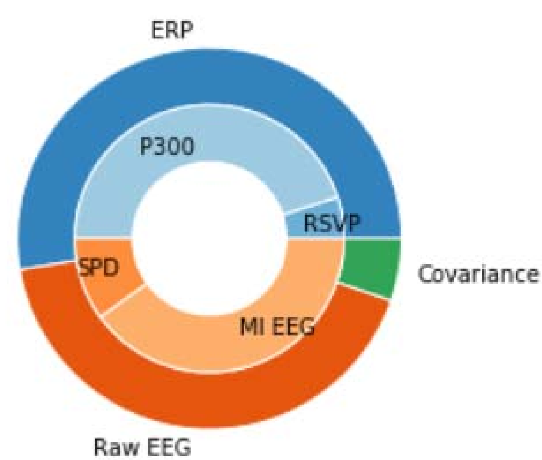


Fig. 1. Usage of MI EEG Data in the Literature.

TABLE. I. CLASSIFICATION OF EEG SIGNALS

Year and Author	Data Type	Signal type	Band-Pass Filter	Feature Extraction	Classification	Result
2019, Zheng wei Wang et al [21]	RSVP-EEG	ERP	5 to 12 Hz	MTWLB	LR	92%
2019, Guan et al [22]	EEG	MI	______	FGMDRM, SJGDA	SSDT	K 0.607
2019, Hao Wu et al [23]	EEG,EOG	MI	Raw	MSFBCNN	MSFBCNN	94.4%
2019, Woo Ha et al [23]	EEG	MI	______	STFT	CapsNet	78.4%
2019, khan et al [24]	EEG	MI		SBCSP-SBFS	SVM,NBPW,KNN	60.61%,86.50%
2018, Hongzhi Qi et al [10]	P300-EEG	ERP	0.1 to 40 Hz.	RG	LDA	90%
2018, Luo et al [27]	EEG	MI	_____	FB-CSP	GRU-RNN with SWCS	82%
2017, Amin et al [28]	EEG	Raw	0-3.90,3.90-7.81	DWT	SVM, KNN,MLP,NB	99,98,97, 89
2017, Oikonomou et al. [13]	MI EEG	Raw	0.5 Hz and 100	CSP , PSD	SVM, LDA	K 0.59, 0.58
2017, Baig et al [12]	MI EEG	Raw	0.05-200 Hz	CSP, DE	LDA,SVM	95%, 95%
2016,Zhang et al [26]	MI EEG	Raw	SBLFB	CSP	Sparse Bayesian learning	81.7%
2016, Ilyas et. al. [14]	EEG	Raw	---	FFT	SVM, K-NN,MLP- ANN,LR	73.03 % 68.97%
2015,Barachant et al [18]	SPD EEG	ERP	Bandpass	--	--	ERP better than Cov
2015,Florian Yger et al [19]	MI-EEG	Cov	8-30 Hz,	CSP	RG	79%
2014, Barachant, Marco [20]	P300-EEG	ERP	1 and 20 Hz.	--	MDM	89%
2014,Gajic et. al [15]	Epileptic EEG rec	raw	173.8 Hz	Wavelet Transform	QDA	99%
2012,Rodríguez et al [25]	EEG	MI	Band pas filter 0.5 to 100 Hz	Power Spectral PSD	SVM,LDA	74.3%

In recent years, MI EEG signals were classified with Subject Specific Decision Tree (SSDT) [22], Multi Scale Filter Bank Convolutional Neural Network (MSFBCNN) [23], Sequential Backward Floating Selection (SBFS) as feature selection and Naïve Bayesian Parzen window (NBPW) used as classification method [24]. Some authors claimed that Power Spectral Density (PSD) is most effective in extracting patterns for classification MI EEG data [13]. Support Vector Machines (SVM) is very famous in EEG classification [14]. SVM with Differential Equation (DE) [12], SVM with PSD [25] showed good performance. Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) are popular for classification of raw EEG signals by attaining accuracy of 99% and 98 % respectively [28]. Sparse Bayesian learning of frequency bands (SBLFB) [26], Recurrent Neural Network (RNN) with Sliding Window Cropping Strategy (SWCS) [27] were also used for MI based BCI. Quadratic Discriminant Analysis (QDA) showed outstanding performance in case of Epileptic EEG data's classification, i.e. 99% [15].

IV. METHODOLOGY

In this section proposed method will be discussed. It revolves around two points. Firstly, find classification accuracy of pipelines before and after implementing Removing Evoked Potential (REP). Secondly, the response of the proposed pipeline, Common Spatial Pattern (CSP) and Gaussian Process Classifier (GPC), in comparison with other pipelines for the classification of those MI EEG signals that have nondiscriminant ERD/ERS values.

A. *Problem Statement*

According to statements of famous researchers

"It is impossible to distinguish between left and right foot motor imagery, or between the movements of particular fingers because the cortical areas associated with these distinct movements are too small to generate Discriminative ERD and ERS signals" [13]. Therefore it is required to study how the classification of these subjects can be improved, which have non-discriminative ERD/ERS signals. This study focuses on those subjects that have non-discriminant ERD/ERS comprising MI EEG signals by using a REP-based filter approach so classification performance can be improved.

B. *Proposed Pipeline*

Common Spatial pattern is one of the best feature extraction methods used for EEG signals. According to research, Gaussian Process Classifier (GPC) is the best classification algorithm. In this study two methods are combined into one pipeline and data after filtering with REP is classified with above mentioned pipeline as well as with state of art algorithms.

a) Common Spatial Pattern: When humans perform some tasks or think about performing some tasks, some signals are generated in the brain. Signals generated while performing something and just imagine of doing something, are somewhat similar. Depending on the type of motor imagery signals, different EEG patterns can be measured. Raw EEG signals have information about signals and noise. It is a linear combination. For feature extraction, Common Spatial

Pattern has obtained good popularity [29]. It finds spatial filters that maximize the ratio of the variance of data of classes [30]. Spatially filtered signal can be described by (1).

$$S = WM \tag{1}$$

Where S is spatially filtered signal, M is Nc x T matrix, W is L x Nc, matrix referred to CSP projection. T is for trial. To understand it in a better way CSP can be visualized as it increases the variance of one class and decreases the variance of other class by transforming axis for data. Therefore, the classes which were not identifiable earlier can be distinguished after using CSP.

b) Gaussian Process Classifier (GPC): Gaussian Process Classifier is used to classify probabilistic classification, it supports multi-class classification by performing either one-versus-rest or one-versus-one based training and prediction. In one-versus-rest [31]. It is a non-parametric classification method and based on Bayesian methodology. It focuses on the modeling of posterior probabilities by defining a certain latent variable fi for the pattern i in case of a two-class problem. To identify whether pattern i belongs to class C1 or not, one will check if fi is positive and large then the probability of belonging pattern i to C1 is high. On the other hand, if fi is negative and large then the probability of belonging to C2 is high. If fi is close to zero then the probability of belonging to any class is ambiguous [13].

C. Proposed Approach

Evoked potentials are computed and averaged for all the trails for every subject. In this case, evoked potentials are averaged for 24 trails. The time duration for calculating is set between -1 and 4 secs. The architecture of the proposed approach is shown in Fig. 2. After computing evoked potentials, it is removed from data while applying band-pass filter. Removing evoked potential is called REP. Two parallel experiments are conducted. One without REP and other after applying REP based filter. Data is visualized to see the effect of REP. Feature extraction and classification is done in parallel. Results are computed and compared to find the performance of REP.

a) REP-based Filter: Frequency filters are used to segregate different frequencies or amplify the desired signal. Frequency filters can be lowpass, highpass, bandpass and bandstop. Cut off frequency is the frequency on which filters start working. Pass filter allows only a particular frequency to pass and stop filter stops particular frequency and allows to pass all others. Bandpass filter allows a certain range of frequencies to pass. Filters are also used to remove noise from signals. Noise in EEG signals can be due to power line noise (50 Hz), eye movement (< 1 Hz), ECG noise (approx 2 Hz), Muscle Unit Potential (70 – 250 Hz).

Requirements of study are, segregation of alpha and beta rhymes, removing noise, removing evoked potential, enhancing low ERD/ERS values in signals. At first, frequency bands are segregated in terms of Alpha and Beta rhymes. Cut off frequency of 7 to 12 Hz is used for Alpha rhyme and 13 to 30 Hz for Beta rhyme. Evoked potentials are calculated and averaged along all the trails. EP values are removed from alpha and beta rhymes. Evoked potentials are removed during bandpass filtering operation on segregated rhymes of data with 53 db stopband attenuation**.** Filtering is done in 3 contiguous segments. A one-pass, zero-phase, non-causal bandpass filter is applied which has a roll of value 6 db. As roll-off value is increased, the sharpness of the filter also increases. Filter length comprises of 529 samples length for time limit 3 sec. Lower transition bandwidth of the filter is set to 1 Hz and upper transition is set to 2 Hz. During filtering evoked potentials are removed. So, data only comprises of ERD/ERS brain's response.

Data can be visualized, before removing evoked potential, from Fig. 3, different colors are showing first 14 channels of data out of 64. Bandpass filter is used to extract frequency bands between 7 to 30 Hz for processing. Resultant signals are enhanced so low ERD/ERS values can be increased. This data is further used in feature extraction and classification step by using the proposed pipeline and other state of art algorithms. The results are computed. In the next step, REP is applied during filter band selection, data after applying REP can be seen from Fig. 4 is showing MI EEG data after filtering with REP approach through bandpass filtering. It can be observed that data is now in a good format after removing evoked potentials. This data is again used in feature extraction and classification step by proposed (CSP + GPC) as well as state of the art algorithm. Results are computed for all the subjects and along all the pipelines.

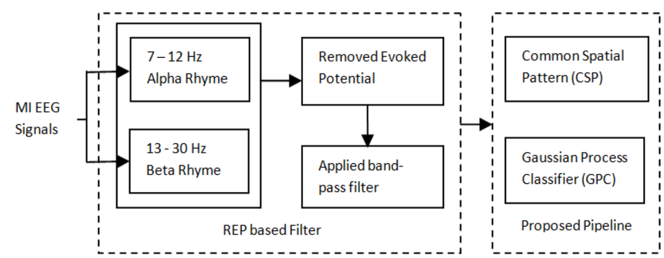


Fig. 2. Architecture of Proposed Method.

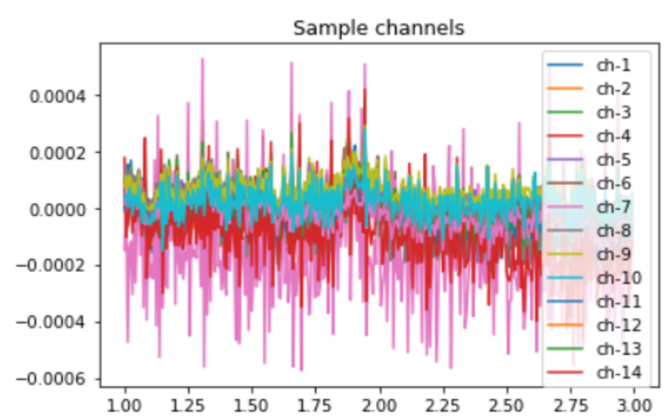


Fig. 3. MI EEG Data before Applying REP.

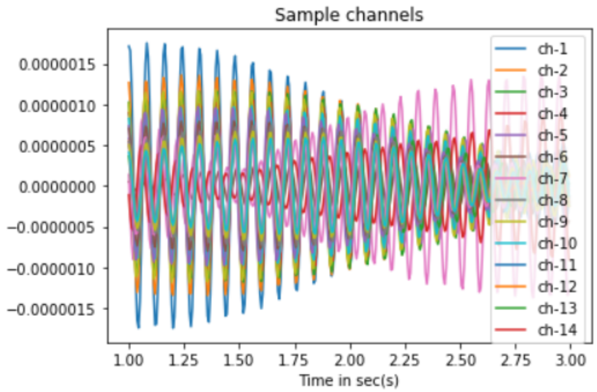


Fig. 4. MI EEG Data after Applying REP.

b) Removing Evoked Potentials: During bandpass filtering evoked potentials are removed from EEG data as shown in Fig. 4. This is done in the preprocessing step while extracting the frequency ranges that are required for the classification of MI EEG signals. The required range is from 7 to 30 Hz which shows frequency bands of Alpha and Beta rhymes as shown in Fig. 5. The sampling frequency used was 160 Hz, yielding 640 samples per sec. Filter lengths used are this study gives 256 and 529 samples per sec for carrying out different experiments. The alpha band is high when a person is in a relaxed state. When a person performs a motor imagery task, a decrease in the alpha frequency band is produced. This decrease is referred to as desynchronization. Desynchronization in Alpha or mu band in this Fig. 5. is showing evidence of the MI task. Alpha and Beta rhymes are showing a there are discriminant ERD and ERS values in this signal. Non-discriminant value of ERD/ERS can be observed from Fig. 6. This figure is showing there is synchronization in an alpha band, but no desynchronization. It shows non-discriminant values or ERD/.ERS. As a decrease in the alpha band shows there is a change in the relaxed mode of the subject in terms of doing some task. There are no discriminant values seen here.

c) Computing ERD/ERS: ERD/ERS is computed using python library MNE. It is a very rich library for processing brain waves. ERD/ERS maps are also visualized as shown in Fig. 7. An ERDS map is a time/frequency representation of ERD/ERS over a range of frequencies. This figure shows a discriminant ERD/ERS map. Blue color shows desynchronization and red color shows synchronization. Cluster-based permutation tests are used for finding significant ERDS values.

d) Computing Common Spatial Patterns: First four components of Common Spatial patterns are computed. CSP is also computed before applying CSP and after applying CSP. All the coding is done in python 3, using the MNE library in anaconda virtual environment. CSP components before applying REP and after applying REP can be seen from Fig. 8 and Fig. 9. After computing CSP components various classifiers are applied. Features are also extracted by using other famous feature extraction method i.e. Xdawn. Several classification algorithms are used for evaluating results on a sound basis. Comparison between the best 7 pipelines is demonstrated in the results section

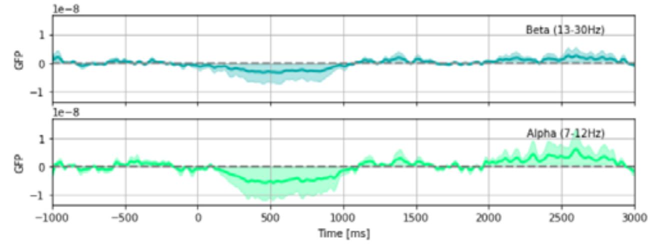


Fig. 5. Alpha Beta Rhyme having Discriminant ERD/ERS.

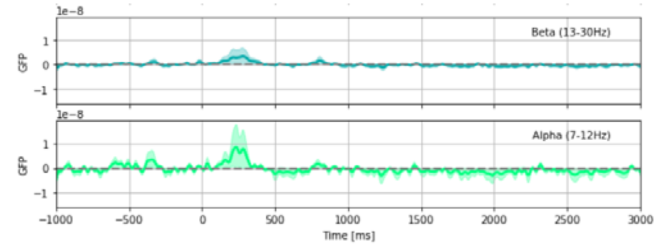


Fig. 6. Alpha Beta Rhyme having Non-discriminant ERD/ERS.

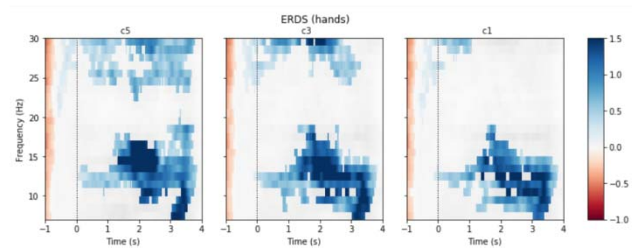


Fig. 7. ERD/ERS Map.

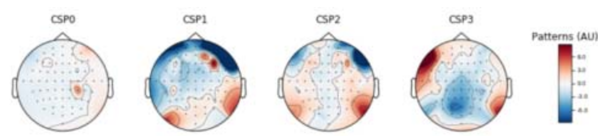


Fig. 8. First Four CSP Components before Applying REP.

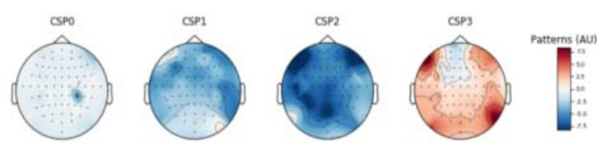


Fig. 9. First Four CSP Components after Applying REP.

V. Results and Discussions

As the main focus of this study is classifying subjects that have low ERD/ERs values. So, in this section, a comparison of classification performance is presented, between the famous seven pipelines for 9 subjects. Results show there is an improvement in the performance of all the classifiers and all the subjects. The performance metric used is ROC-AUC. That is a famous performance metric and used in lots of research studies.

a) Datasets: Data is taken from an online data resource, physioNet[1]. Dataset was recorded from 64-channels EEG using the BCI2000 system. 14 runs of experiments were performed by each subject. Two one-minute baseline runs

[1] https://physionet.org/about/database/

(one with eyes open, one with eyes closed), and three two-minute runs of each of the four following tasks as shown in Table II.

There are in total 14 runs of experiments performed by each subject. Data is in EDF+ format. It contains 64 channels for each signal and sampled at 160 samples per sec. each subject performed 46.62 ± 0.96 trials for the left- and right-hand motor imagery tasks. The average numbers of trials were 23.62 ± 0.61 and 23.00 ± 0.62 for the left- and right-hand motor imagery data respectively. The EEG data is sampled at 160 Hz for all subjects yielding 640 samples for every single trial.

b) Experiments: All the experiments are done by using the intel core i7 processor having 8 GB RAM and 64-bit operating system. Coding is done using python 3, in anaconda platform. A virtual environment is set by importing all the libraries in it. Namely mne, sklearn, matplotlib, numpy and pyriemann. Dataset is taken from physioNet, an online data recourse. Data is taken only for the required runs, that are associated with motor imagery tasks.

The performance of the proposed pipeline is evaluated by using two types of datasets, one without using REP approach and other after applying REP. Same course is repeated for other pipelines including state of art algorithms. Experiments are conducted and visualized by using all the above libraries in jupyter notebook.

Here are actual screenshots of results from code for comparing results before applying REP and after applying REP from Fig. 10 – Fig 18. Classification pipelines that are used in this study are, Vectorizer with Gaussian Process Classifier (Vect + GPC), Common Spatial Pattern with Gaussian Process Classifier (CSP + GPC), Common Spatial Pattern with Support Vector machines (CSP + SVM, State of art algorithm), Common Spatial Pattern with Quadratic Discriminant Analysis (CSP + QDA, State of art algorithm), Common Spatial pattern with Logistic Regression (CSP + LR, State of art algorithm) and XDawn with GPC. The Second Dataset is taken from Github. Data is acquired from 32 channel electrodes, there are in total of 34 columns and 3469302 rows. There are 3 event types. Each sample will run for 1300 ms Data is preprocessed to remove Null values. After that REP based filter is applied. The results of both datasets are given in the following section.

c) Results: This section shows the results of various pipelines for every subject, before and after applying REP. Table III is showing the classification accuracies of all the pipelines used. It can be observed that the performance of all pipelines is improved.

Fig. 10 to Fig. 18 shows classification performance of pipelines before applying REP and after using REP. The performance of pipelines for each subject is demonstrated separately, to show the precision of results.

Fig. 20 is showing that the performance of all the pipelines is improved by using REP. Vect + GPC is showing the maximum increase in performance. Classification performance is improved significantly by all the pipelines except Xdawn + GPC. The performance of CSP + QDA for subject 7 improved remarkably. The focus of this study is upon improving the classification of subjects having low-ERD/ERS values. Fig. 19 shows the classification performance of all the subjects has increased after applying REP. The Minimum increase in performance metric is 4% and the maximum increase is 14%. An increase in the performance of classifiers makes a system accurate. Mean accuracies of pipelines before and after application of REP is given in Table III. Table II demonstrates that the classification accuracy of all the pipelines which are used in the experiment also increased. The graphical representation of this table is shown in Fig. 20. To be more specific, the percentage increase in pipelines can be seen by Fig. 21. The minimum increase is 1% and the maximum increase is 20%. From the above results, it is obvious that REP is a robust approach and can be used for increasing the performance of state of the art algorithms. The results of the second dataset are as shown in Table IV and Fig. 22.

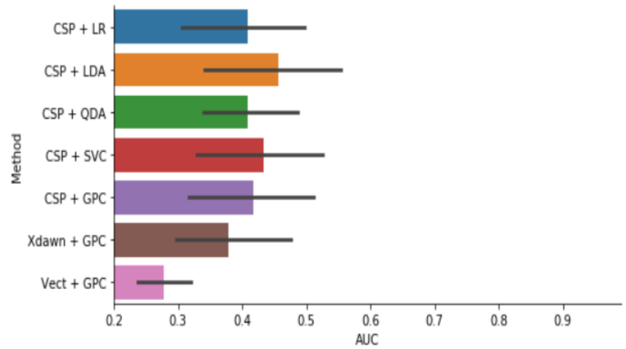
Table IV shows the results of the second dataset of only one pipeline, i.e. CSP + GPC. The graphical representation of results is shown in Fig. 22. To prove the robustness of the proposed method one more filter "Infinite Impulse Response (IIR)" is also applied to MI EEG data. Given are the results of three filters, namely, Finite Impulse Response (FIR), Infinite Impulse Response (IIR) and Removed Evoked Potential (REP). Results (by using three filters) of one subject are demonstrated by Fig. 23. The mean values for all the subjects are shown in Fig. 24.

TABLE. II. TASKS PERFORMED BY SUBJECTS

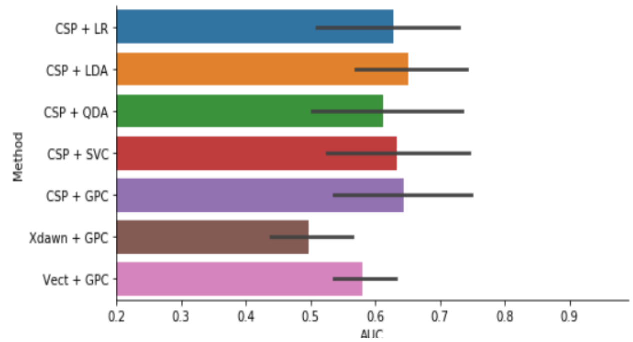
Task no.	Target Location	Task
1	Left side of screen	Subject opens or closes left fist
2	Right side of screen	Subject opens or closes right fist
3	Top of screen	Opens or closes both fist and right foot
4	Bottom of screen	Open or closes both fist and left foot

TABLE. III. MEAN ACCURACIES OF ALL THE PIPELINES

Pipelines3	Not REP	REP
Vect + GPC	40%	60%
CSP + GPC	54%	69%
CSP + SVM	49%	65%
CSP + QDA	45%	64%
CSP + LDA	50%	68%
CSP + LR	51%	65%
Xdawn + GPC	48%	49%

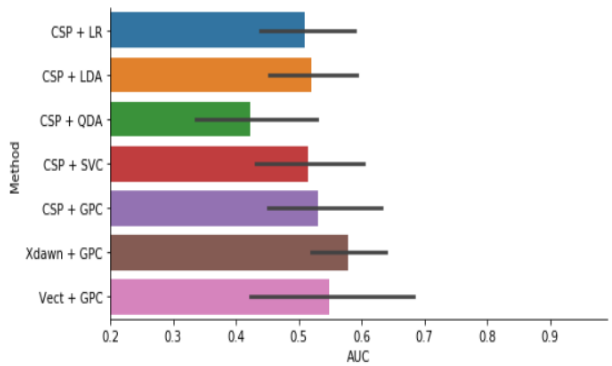


Classification Result before REP

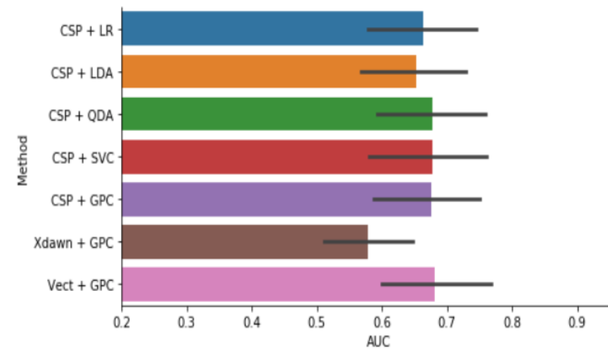


Classification Result after REP

Fig. 10. Classification Result of Seven Pipelines for Subject 1.

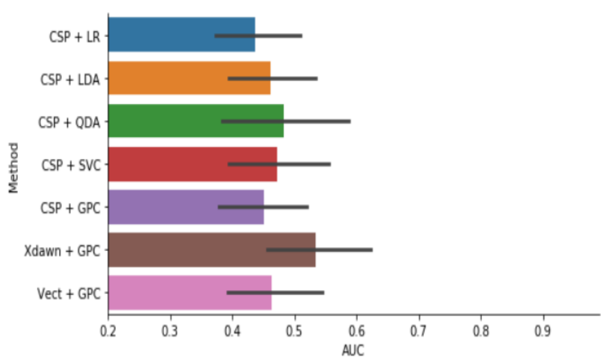


Classification Result before REP

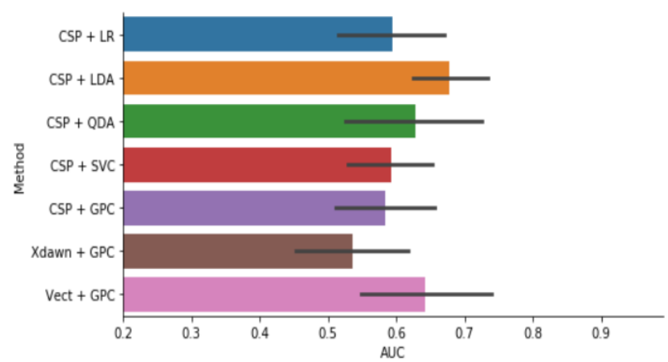


Classification Result after REP

Fig. 11. Classification Result of Seven Pipelines for Subject 2.

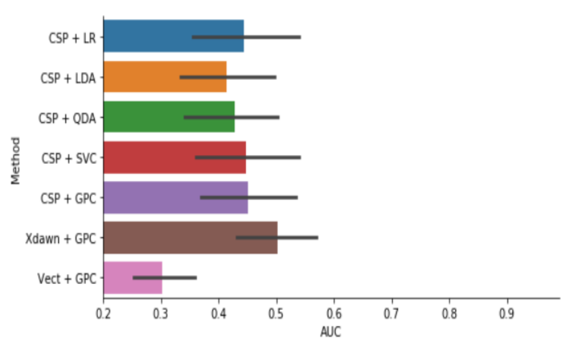


Classification Result before REP

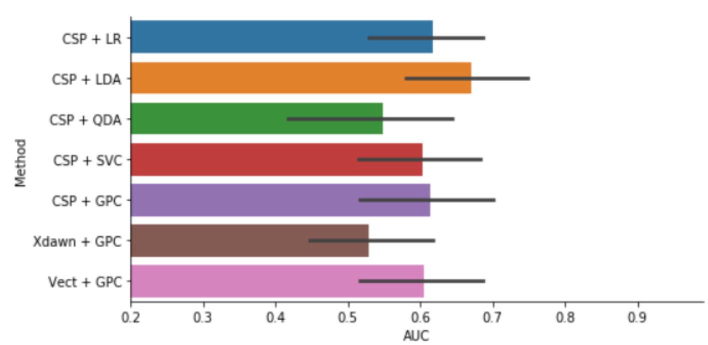


Classification Result after REP

Fig. 12. Classification Result of Seven Pipelines for Subject 3.

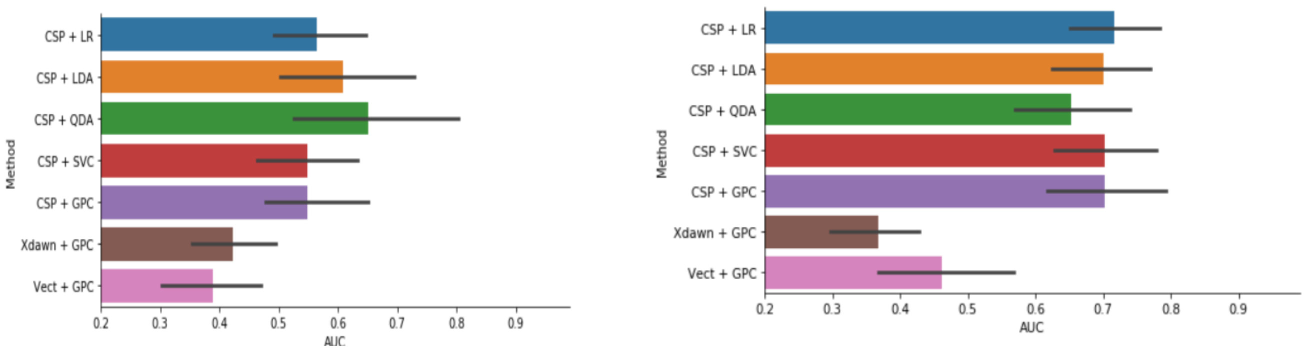


Classification Result before REP



Classification Result after REP

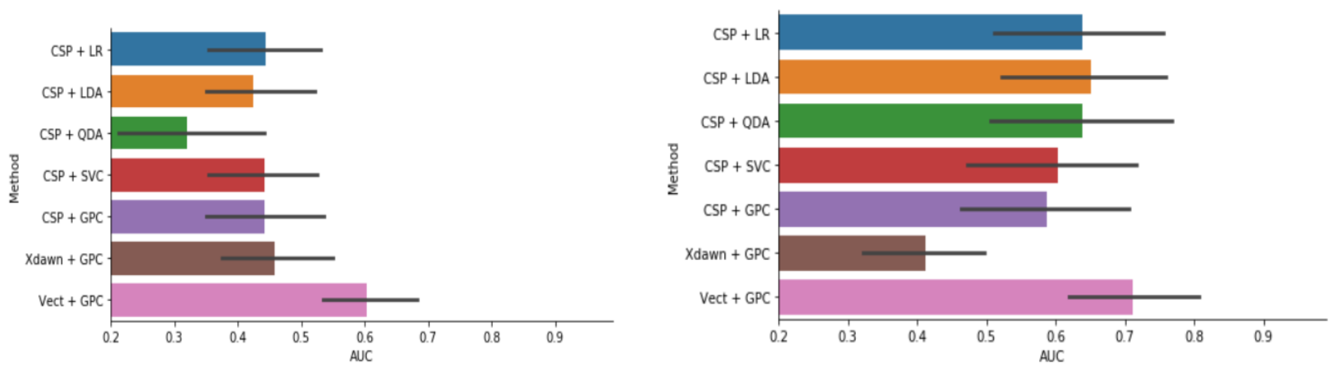
Fig. 13. Classification Result of Seven Pipelines for Subject 4.



Classification Result before REP

Classification Result after REP

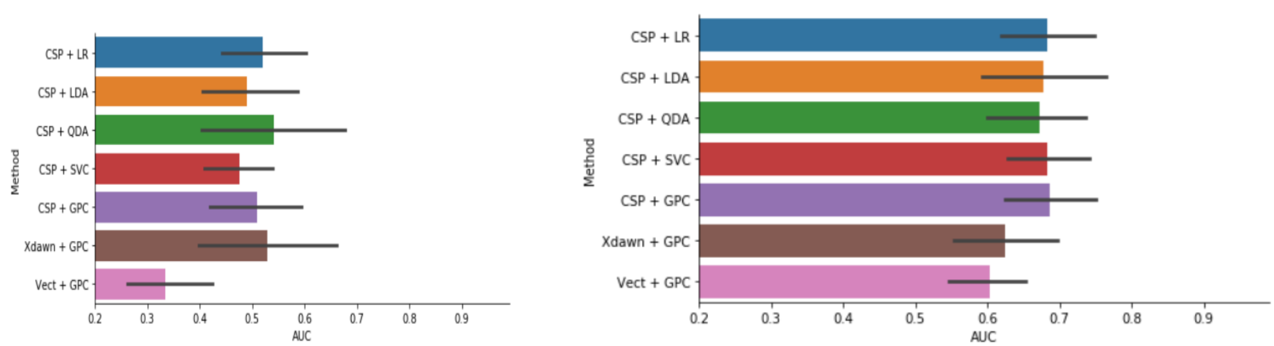
Fig. 14. Is Showing Classification Result of Seven Pipelines for Subject 5.



Classification Result before REP

Classification Result after REP

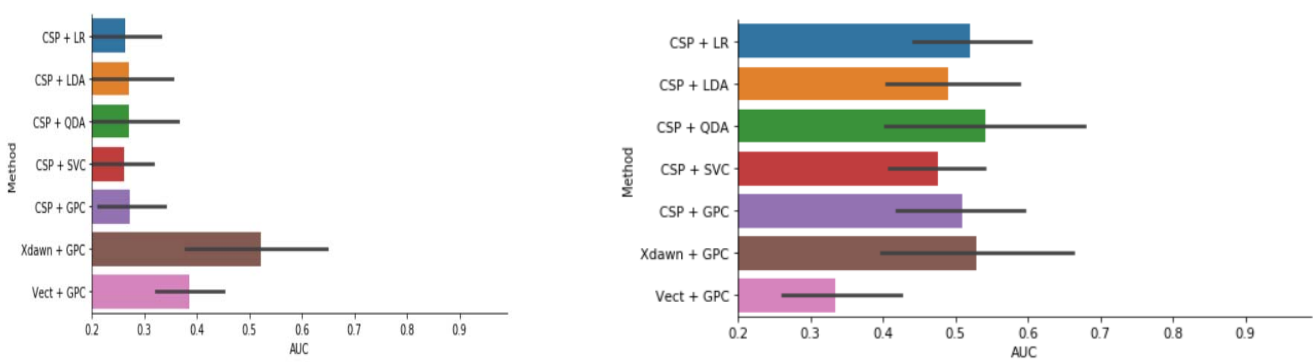
Fig. 15. Is Showing Classification Result of Seven Pipelines for Subject 6.



Classification Result before REP

Classification Result after REP

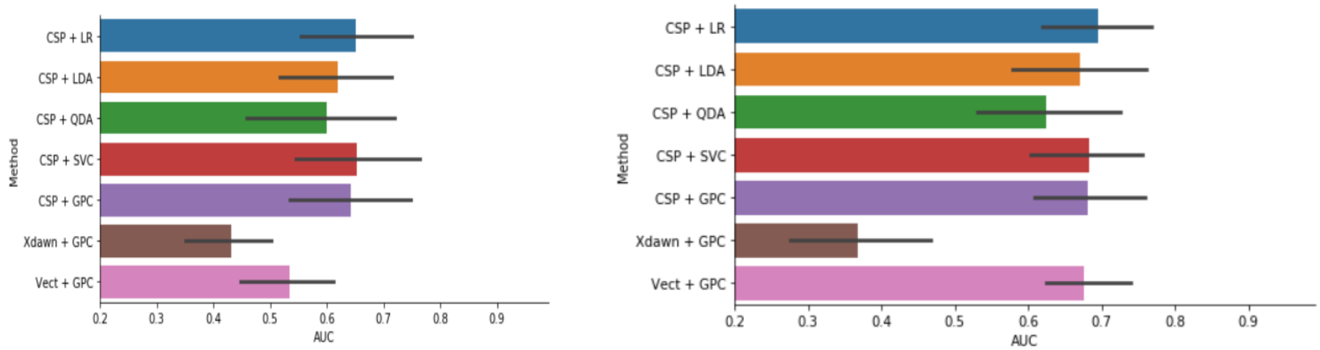
Fig. 16. Is Showing Classification Result of Seven Pipelines for Subject 7.



Classification Result before REP

Classification Result after REP

Fig. 17. Is Showing Classification Result of Seven Pipelines for Subject 8.



Classification Result before REP

Classification Result after REP

Fig. 18. Is Showing Classification Result of Seven Pipelines for Subject 9.

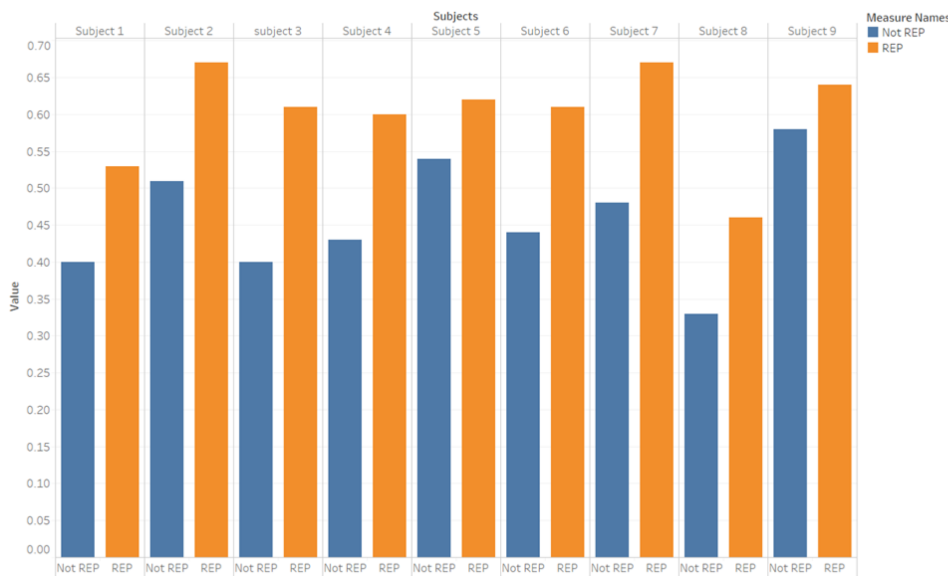


Fig. 19. Increase in Classification Accuracy of all the Subjects.

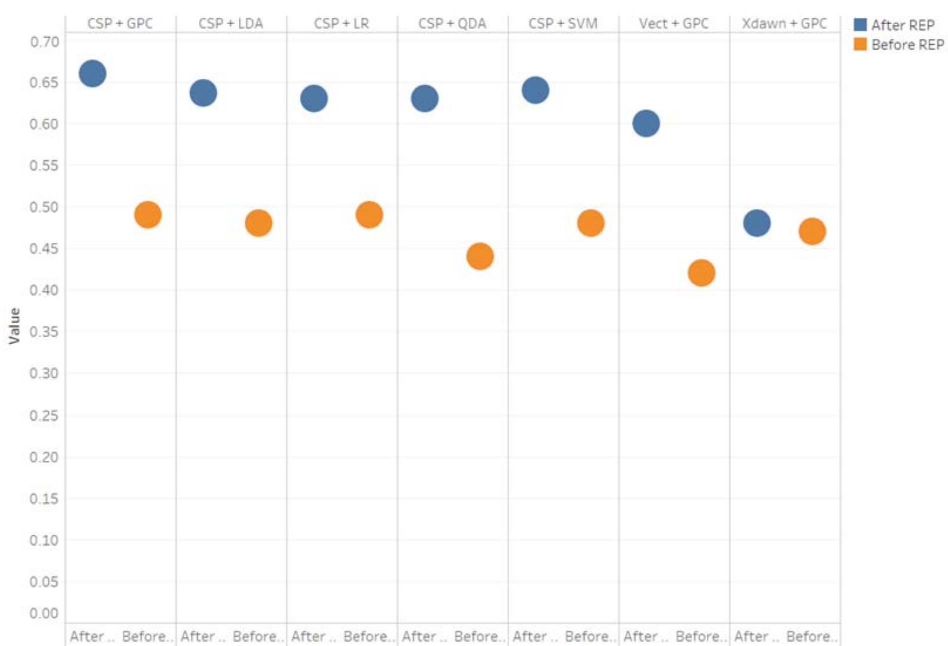


Fig. 20. Increase Classification Accuracy of all the Pipelines, this Figure shows the Performance of CSP + GPC is better than all other Pipelines.

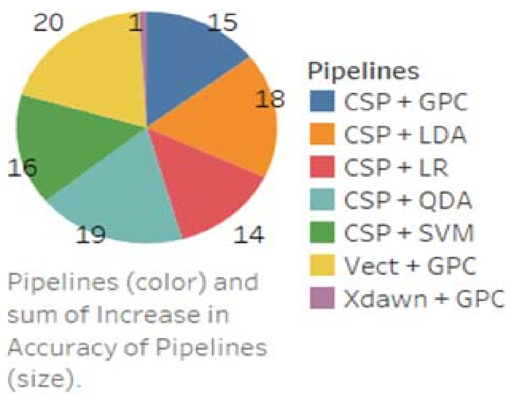


Fig. 21. Increase in Classification Accuracy of all the Pipelines.

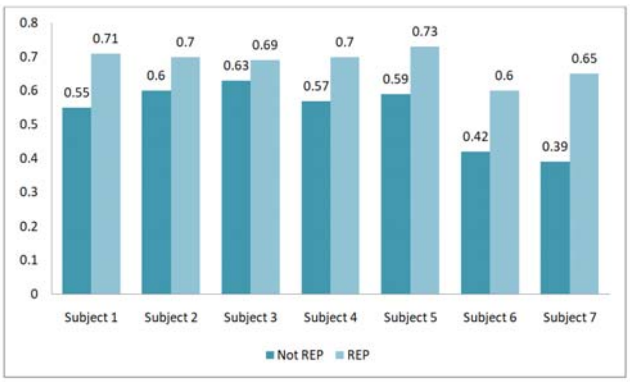


Fig. 22. Classification Result of CSP + GPC for Second Dataset.

TABLE. IV. ACCURACIES OF PROPOSED PIPELINE FOR SECOND DATASET

Subjects 3	Not REP	REP
1	0.55 ± 0.05	0.71 ± 0.01
2	0.60 ± 0.07	0.70 ± 0.05
3	0.63 ± 0.16	0.69 ± 0.08
4	0.57 ± 0.04	0.70 ± 0.11
5	0.59 ± 0.05	0.73 ± 0.01
6	0.42 ± 0.01	0.60 ± 0.05
7	0.39 ± 0.06	0.65 ± 0.03

For showing the robustness of the proposed approach, the results of REP are compared with two other filters. Fig. 24 is showing mean accuracies of all the pipelines used in this study.

d) Discussion: In datasets, some subjects did MI tasks with ease, while some others faced difficulty. This affects the results of the classification in the case of MI EEG signals. Before applying REP the classification accuracy of all subjects, by most of the pipelines, was not satisfactory. After using REP in filter band selection, the performance of all classifiers improved. Some pipelines performed better in the case of one subject whereas others performed well for other subjects, but in most cases, their AUC value was approximately equal to 0.50. REP approach improved performance by increasing AUC value to around 0.70. Less than 0.50 shows classifier is performing reversely, that is saying label 0 as 1 and label 1 as 0. AUC value equals 0.50 means classifiers are unable to detect a class of data. AUC value around 0.70 means classifiers is performing well. Similarly, AUC value around 0.90 means the classifier is performing excellently.

For subject 1, CSP + LR got AUC value 0.42± 0.10 before applying REP and 0.63±0.10 after applying REP. For subject 2, CSP + QDA got AUC value 0.42 ± 0.10 before applying REP and 0.69 ± 0.05 after applying REP. For subject 3, CSP + LDA attained AUC value 0.46 ± 0.11 before applying REP and 0.69 ± 0.07 after REP. It gained maximum increases among all pipelines for subject 3, i.e. 0.23. Minimum increase is attained by Xdawn + GPC pipeline, i.e. is 0.001. Different colors are showing different pipelines. Blackline is showing an interval of confidence for each pipeline. Vect + GPC pipeline also showed good result after applying REP, it had AUC value 0.46 ± 0.10 before REP and gained 0.63 + 0.10 after REP. The gain in AUC value is 0.17. For subject 4, Vect + GPC got AUC value 0.30 ± 0.05 before applying REP and 0.60 ± 0.80 after applying REP. It gained maximum increases among all pipelines. Subject 5 has these results, CSP + GPC and CSP + SVC got AUC value 0.50 ± 0.10 and 0.50 ± 0.10 respectively before applying REP and 0.70 ± 0.12 and 0.70 ± 0.09 after using REP. They gained maximum increases among all pipelines, i.e. 0.20.

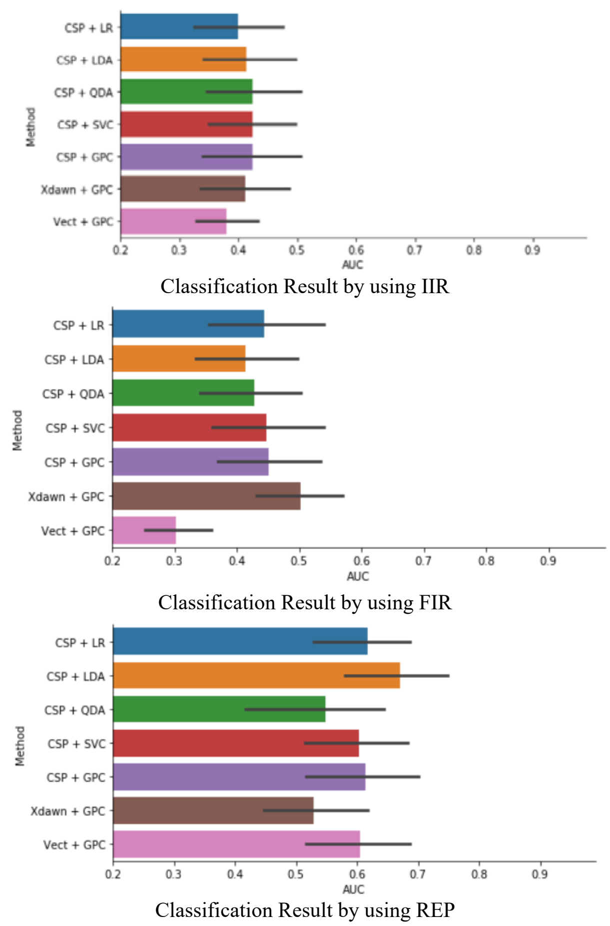


Fig. 23. Classification of Same Subject by Three different Filters.

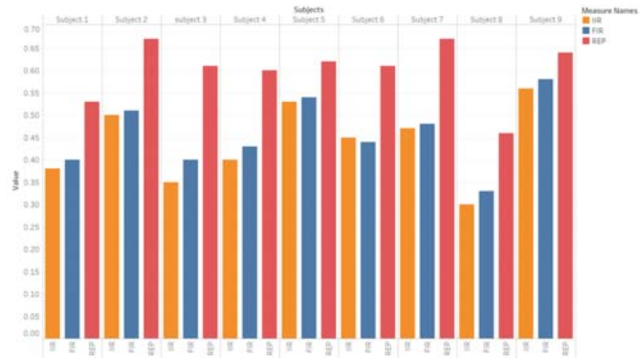


Fig. 24. Comparison of Results of Three Filters.

For subject 6, CSP + QDA got AUC value 0.32 ± 0.10 before applying REP and 0.62 ± 0.16 after applying REP. They gained maximum increases among all pipelines, i.e. 0.30. Auc value of Xdawn + GPC decreased instead of increasing, i.e. it went down from 0.46 ± 0.10 to 0.40 ± 0.10. Subject 7 has these results; CSP + SVC got AUC value 0.48 ± 0.04 before applying REP and 0.68 ± 0.07 after applying REP. They gained maximum increases among all pipelines, i.e. 0.20. Subject 8 results' are, CSP + QDA got AUC value 0.30 ± 0.1 before applying REP and 0.55 ± 0.07 after applying REP. They gained maximum increases among all pipelines, i.e. 0.25. Subject 9 has these values, Vect + GPC has AUC value 0.51 ± 0.09 before applying REP and 0.68 ± 0.05 after applying REP. They gained maximum increases among all pipelines for Subject 9, i.e. 0.17.

The proposed approach increased the performance of classifiers for low ERD/ERS subjects from null to good. Fig. 19 is showing there is an improvement in the classification of every subject. Three experiments are conducted in parallel. In the first one, the classification accuracy of MI EEG signals is found by using a FIR filter. In the second one, the classification accuracy of MI EEG signals is observed by using an IIR filter. In the third one, a REP-based filter is used. Their results are compared as sown in Fig. 10 to Fig. 24. It is found the performance of most of the classifiers improved. Some classification pipelines didn't show any significant improvement i.e. Xdawn comprising pipelines (because it deals with evoked potentials). Therefore on removing evoked potentials from MI EEG data, its classification accuracy decreased in some cases, i.e. Fig. 14 and Fig. 15. In most cases, the performance of the Xdawn comprising pipeline remained the same before and after applying REP. All the pipelines that used CSP as a feature extraction technique were affected by the REP approach in terms of an increase in performance. It shows the performance of CSP comprising pipelines improved by REP.

VI. Conclusion

This paper proposes an efficient REP based filter approach for the classification of MI EEG signals. The proposed method employs a REP based filter approach with CSP as a feature extraction technique and GPC as a classification method. In the proposed method REP based filter is applied on MI EEG signals to remove evoked potential. The remaining signal comprises only one response i.e. ERD/ERS. Proposed Pipeline (CSP + GCP) and various other the state of the art algorithms are applied on this REP based filtered data as well as on data where REP based approach not used. Results of comparison show a clear improvement in classification accuracy of state of the art algorithms as well as in proposed one i.e. up to 20%. Results of the second data set are also showing improvement after the application of REP based filter. As future work REP based approach will be applied on a self-acquired MI EEG dataset to implement it in the real-time scenario.

References

[1] F. Yger, M. Berar, and F. Lotte, "Riemannian Approaches in Brain-Computer Interfaces: A Review," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 10, pp. 1753–1762, 2017.

[2] G. F. Woodman, "NIH Public Access," vol. 72, no. 8, pp. 1–29, 2013.

[3] N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren, "Motor-Imagery : Techniques and Challenges," pp. 1–34.

[4] G. Pfurtscheller and F. H. Lopes, "Event-related EEG / MEG synchronization and desynchronization : basic principles," vol. 110, pp. 1842–1857, 1999.

[5] H. I. Virus, A. N. Disorders, C. Report, M. C. Author, and A. C. Author, "ce pte d M an us cri Ac ce pte d M us pt," pp. 1–46, 2012.

[6] F. Lotte et al., "A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces: A 10-year Update," J. Neural Eng., pp. 0–20, 2018.

[7] H. Xie, D. Xiao, and B. Xia, "The Research for the correlation between ERD / ERS and CSP," pp. 1872–1876, 2011.

[8] U. E. Control et al., "A Brain-Machine Interface Based on ERD/ERS for an Upper-Limb Exoskeleton Control," pp. 1–14, 2016.

[9] H. Cho, M. Ahn, S. Ahn, M. Kwon, and S. Chan, "EEG datasets for motor imagery brain computer interface."

[10] P. Speller, "A Speedy Calibration Method Using Riemannian Geometry Measurement and Other-Subject Samples on A," vol. 4320, no. 91648122, 2018.

[11] J. Kevric and A. Subasi, "Biomedical Signal Processing and Control Comparison of signal decomposition methods in classification of EEG signals for motor-imagery BCI system," Biomed. Signal Process. Control, vol. 31, pp. 398–406, 2017.

[12] M. Z. Baig, N. Aslam, H. P. H. Shum, and L. Zhang, "Differential evolution algorithm as a tool for optimal feature subset selection in motor imagery EEG," Expert Syst. Appl., vol. 90, pp. 184–195, 2017.

[13] V. P. Oikonomou, K. Georgiadis, G. Liaros, S. Nikolopoulos, and I. Kompatsiaris, "A Comparison Study on EEG Signal Processing Techniques Using Motor Imagery EEG Data," Proc. - IEEE Symp. Comput. Med. Syst., vol. 2017-June, no. 1, pp. 781–786, 2017.

[14] M. Z. Ilyas, P. Saad, and M. I. Ahmad, "Classification of EEG Signals for Brain-Computer Interface Applications : Performance Comparison *," 2015.

[15] D. Gajic, Z. Djurovic, S. Di Gennaro, and F. Gustafsson, "Classification of Eeg Signals for Detection of Epileptic Seizures based on Wavelets And Statistical," vol. 26, no. 2, pp. 1–13, 2014.

[16] J. Zhou, M. Meng, Y. Gao, Y. Ma, and Q. Zhang, "Classification of Motor Imagery EEG Using Wavelet Envelope Analysis and LSTM Networks," 2018 Chinese Control Decis. Conf., pp. 5600–5605, 2018.

[17] A. Barachant, "Information Geometry : A framework for manipulation and classification of Neural Timeseries," 2017.

[18] M. Congedo and A. Barachant, "A special form of SPD covariance matrix for interpretation and visualization of data manipulated with Riemannian geometry," AIP Conf. Proc., vol. 1641, pp. 495–503, 2015.

[19] F. Yger et al., "Averaging Covariance Matrices for EEG Signal Classification based on the CSP : an Empirical Study To cite this version : HAL Id : hal-01182728," 2015.

[20] A. Barachant and M. Congedo, "A Plug & Play P300 BCI Using Information Geometry," pp. 1–9.

[21] Z. Wang, G. Healy, A. F. Smeaton, T. E. Ward, and I. V Jan, "ORIGINAL RESEARCH Spatial Filtering Pipeline Evaluation of

Cortically Coupled Computer Vision System for Rapid Serial Visual Presentation," 2019.

[22] S. Guan, K. Zhao, and S. Yang, "Motor Imagery EEG Classification Based on Decision Tree Framework and Riemannian Geometry," Comput. Intell. Neurosci., vol. 2019, 2019.

[23] H. Wu et al., "A Parallel Multiscale Filter Bank Convolutional Neural Networks for Motor Imagery EEG Classification," vol. 13, no. November, pp. 1–9, 2019.

[24] J. Khan, M. H. Bhatti, U. G. Khan, and R. Iqbal, "Multiclass EEG motor-imagery classification with sub-band common spatial patterns," Eurasip J. Wirel. Commun. Netw., vol. 2019, no. 1, 2019.

[25] G. Rodŕiguez-Beŕmudez and P. J. Gárcia-Laencina, "Automatic and adaptive classification of electroencephalographic signals for brain computer interfaces," J. Med. Syst., vol. 36, no. SUPPL.1, 2012.

[26] Y. Zhang, Y. Wang, J. Jin, and X. Wang, "Sparse Bayesian learning for obtaining sparsity of EEG frequency bands based feature vectors in motor imagery classification," Int. J. Neural Syst., vol. 27, no. 2, pp. 1–13, 2017.

[27] T. jian Luo, C. le Zhou, and F. Chao, "Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network," BMC Bioinformatics, vol. 19, no. 1, pp. 1–18, 2018.

[28] H. U. Amin, W. Mumtaz, A. R. Subhani, M. N. M. Saad, and A. S. Malik, "Classification of EEG signals based on pattern recognition approach," Front. Comput. Neurosci., vol. 11, no. November, pp. 1–12, 2017.

[29] J. Wang and Z. Feng, "Feature Extraction by Common Spatial Pattern in Frequency Domain for Motor Imagery Tasks Classification," pp. 5883–5888, 2017.

[30] I. Xygonakis, A. Athanasiou, N. Pandria, D. Kugiumtzis, and P. D. Bamidis, "Decoding Motor Imagery through Common Spatial Pattern Filters at the EEG Source Space," vol. 2018, no. Mi, 2018.

[31] M. Opper and O. Winther, "Gaussian processes for classification: Mean-field algorithms," Neural Comput., vol. 12, no. 11, pp. 2655–2684, 2000.

Barchan Sand Dunes Collisions Detection in High Resolution Satellite Images based on Image Clustering and Transfer Learning

M. A. Azzaoui[1], L. Masmoudi[2]
Laboratoire Conception et Systèmes: Électronique, signaux et informatique
Université Mohammed V de Rabat
Morocco

H. El Belrhiti[3]
Département des Sciences Fondamentales et Appliquées
Institut Agronomique et Vétérinaire Hassan II, Rabat, Morocco

I. E. Chaouki[4]
Département Informatique
Ecole Nationale des Sciences Appliquées d'Agadir
Agadir, Morocco

***Abstract*—Desertification is a core concern for populations living in arid and semi-arid areas. Specifically, barchans dunes which are the fastest moving sand dunes put constant pressure over human settlements and infrastructure. Remote sensing was used to analyze sand dunes around Tarfaya city located in the south of Morocco in the Sahara Desert. In this area, dunes form long corridors made of thousands crescent shaped dunes moving simultaneously, thus, making data gathering in the field very difficult. A computer vision approach based on machine learning was proposed to automate the detection of barchans sand dunes and monitor their complex interactions. An IKONOS high resolution satellite image was used with the combination of a clustering algorithm for image segmentation of the dunes corridor, and a Transfer Learning model which was trained to detect three classes of objects: Barchan dunes, bare fields and a new introduced class consisting of dunes collisions. Indeed, collisions were very difficult to model using classical digital image processing methods due to the large variability of their shapes. The model was trained on 1000 image patches which were annotated then augmented to generate a larger dataset. The obtained detection results showed an accuracy of 84,01%. The interest of this research was to provide with a relatively affordable approach for tracking sand dunes locations in order to better understand their dynamics.**

Keywords— High resolution satellite images; remote sensing; transfer learning; image segmentation; sand dunes; desertification

I. Introduction

A. Barchan Dunes and Desertification Issue

The United Nations declared desertification as a phenomenon affecting 250 million people worldwide and increasing in intensity as a result of climate change [1]. Not all dunes contribute equally to the desertification problem. Indeed, the taxonomy of sand dunes includes multiple dune categories, posing different risk levels depending on their velocities [2]. The 3 main categories of sand dunes are: Star dunes, Linear dunes and Barchan dunes. The first ones called Star Dunes form in areas where winds blow from multiple directions, thus limiting their moving capabilities. The second ones called Linear Dunes form under bi-directional winds that define their longitudinal shape, thus making them relatively slow. The third ones called Barchans form under unidirectional winds, thus, they are the fastest moving sand dunes, and have characteristic crescent shape [3]. In the context of natural hazard monitoring, data gathering is compulsory to anticipate barchans dunes risk. However, it is obviously expensive and difficult to gather field data regularly at such a large scale in a harsh environment. Hence, the use of remote sensing, which was a convenient solution.

B. Problem Statement

Remote sensing is essential for monitoring sand dunes, however, as deserts contain thousands of dunes; and their annotation is executed manually, this task becomes time consuming. Therefore, the proposed approach consisted of investigating the use of computer vision algorithms to automate barchans dunes detection and segmentation. Barchan dunes had a characteristic crescent shape, which could be segmented using classical computer vision algorithms [4]. However, in the field, thousands of barchans moved simultaneously, with smaller dunes moving faster than larger ones, resulting into dune collisions. This further complicates the problem, as dunes collisions had odd shapes which require more advanced computer vision algorithms to be detected. Therefore, in this study, Transfer Learning was introduced to detect barchans dunes collisions.

C. Aim of Research

The proposed research study allowed to gather data through the use of high resolution satellite imagery along with a machine learning approach. The goal was to provide with valuable information, not only to geologists, which would to improve their understanding of desertification phenomenon, but also, to urban planners towards the selection of sustainable solutions for reducing the effect of moving sand dunes on the daily lives of the local inhabitants and all stakeholders, and also for decision makers, to better anticipate the risks of moving sand dunes and limit desertification phenomenon impact on the populations.

D. Sand Dunes Detection using Remote Sensing

Sand dunes detection is an example of application for which the use remote sensing was essential. In fact, scientists started using it early on for the categorization of sand dunes: [5] used Landsat (ERTS) mosaics to create a global map of sand 'seas'. [6] used QuickBird satellite images to observe morphometric changes in sand dunes. Aerial photogrammetry was also common: [7] analysed the self-organizing patterns of sand dunes. [8] used Google maps images to confirm simulations for dunes stabilization. However, dunes mapping was mostly done manually. With the development of digital image processing, the use of satellite images became more efficient and more information could be extracted from images. During this phase, pixel based approaches were preponderant in the literature [9], [10], [11], and spatial and spectral properties of sand dunes were extensively exploited to gather morphology measurements and study their dynamics, based on ASTER, IKONOS or Landsat satellite images. Later on, with the development of computer vision, researchers in remote sensing shifted to Geospatial Object Based Image Analysis (GEOBIA) paradigm, and dunes detection approaches were based on objects characteristics such as shape, motion, colour and texture [12]. With the development of image descriptors such as (GLCM) Gray-Level Co-occurrence Matrix for texture, Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP) and Speed-Up Robust Features (SURF), more approaches were focused on feature engineering, and the use of classifiers with ensemble classifiers such as random forest (RF), and as support vector machine (SVM) to achieve dunes detections results. In recent years, the advances made in machine learning and especially Neural Networks with Deep Learning became more popular in remote sensing.

E. Machine Learning and Transfer Learning for Satellite Remote Sensing

Machine Learning includes many training algorithms and models, among which Support Vector Machines (SVM) and Multi-Layer Perceptrons (MLP) which could be used to solve non-linear problems. Indeed, in the case of (MLP), the model is composed by a set of hidden layers which learn successively higher level features, in addition to an input layer and an output layer rendering the classification [13]. Layers were made of neurons which were characterized by weight coefficients w_i, and an activation function ϕ. When the number of hidden layers is important, the model gets called Deep Learning [14]. Deep Learning was also trained using Backpropagation algorithm which dates back to the 90s. During the 2000s, other machine learning algorithms and models such as (SVM) were outperforming (MLPs), and adding layers was considered as an explosion of free parameters, producing a high VC dimension, which would result into overfitting. The turning point came on 2012 when AlexNet, a Convolutional Neural Network (CNN) won the ImageNet competition by a significant margin, by combining the use of a large dataset, rectified linear units (ReLUs) as activation functions, and graphics processing units (GPUs). Computer vision community gained interest and other classes of Deep Learning models were re-explored and developed: Autoencoders (AE) which were used for unsupervised learning, Deep Belief Networks (DBN), and also Generative Adversarial Networks (GAN) models, where two neural networks competed with each other as in game theory [15]. Other developments took place in the recent years with VGG Networks [16], RESNET [17] and Inception-v4 [18].

Remote sensing researchers have used Deep Learning in different applications: [19] for hyperspectral data classification, [20] reviewed Deep Learning for in Land Use and Land Cover with applications related to urbanization. [21] showed the use of Deep Learning for image processing, object detection and classification, including aircraft segmentation, vehicle detection, and scene identification. However, to the extent of our knowledge, no study applied Deep Learning and its variant called Transfer Learning which will be presented in detail in the next section, for the detection of Barchan dunes and sand dunes collisions in high resolution satellite images.

F. Our Contribution

The former paragraphs exposed the importance of barchan dunes detection for treating the desertification issue. They also showed how dunes mapping required remote sensing, and explained that the current approaches for mapping sand dunes were manual and occasionally, made use of basic digital image processing, which limited their capabilities to objects with well-defined shapes. This is where this contribution intervenes: As existing approaches could not generalize well to the variability of dunes shapes and especially sand dunes collisions, this paper introduced a new approach that combined Transfer Learning which is based on Deep Learning, and image clustering for the detection of Barchan dunes and their collisions in remote sensing satellite images. In the next sections, the framework used will be presented in details, along with the obtained results, a discussion of future works, and the conclusion.

II. Proposed Work

The proposed method was composed of three major steps: First, it began with a pre-processing step on the satellite image, in order to increase the spatial resolution of multispectral bands, reduce noise and enhance image contrast. Second, it used an image clustering method to segment sand regions and distinguish them from the bedrock. Third, it trained a Transfer Learning framework on areas which were segmented previously, thus combining both methods to predict three classes of objects in the satellite image: barchans dunes, dunes collisions and bare field. In the following, first, a brief background presentation is presented for Transfer Learning, which is a derivation of a Deep Learning model. Next, the different steps of the proposed method are explained in detail.

A. Transfer / Deep Learning Background

Transfer Learning is an approach used to bootstrap the learning of a new task, by transferring knowledge from a similar task already learned, and adapting it. In our case, it was Deep Learning AlexNet architecture. AlexNet is a Deep Learning Convolutional Neural Network (CNN) which was first introduced during ImageNet contest where it outperformed state of art methods by 10% [22]. Its structure consisted of 8 layers: 3 connected layers and 5 convolutional layers. illustrates its architecture which was composed of 60 million parameters and 650,000 neurons in total. This model took 3 channels RGB images of size 227×227, which went through its consecutive layers (Fig.1).

The first layer was a convolutional layer with 96 kernels having a size of 11x11x3 and a ReLU non-linear activation function (1).

$$\emptyset(x) = \max(0, x) \quad (1)$$

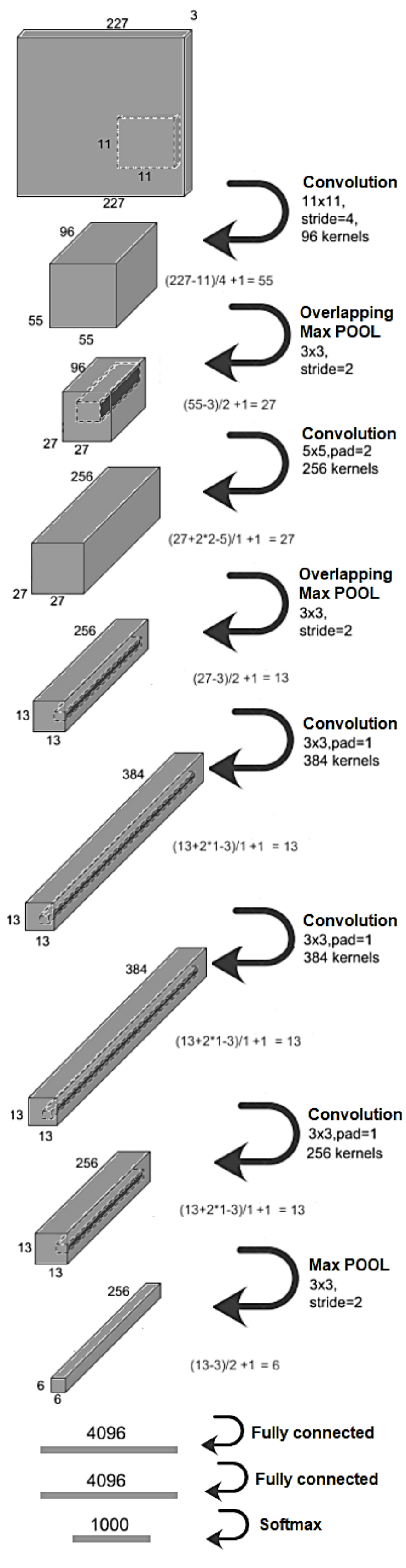


Fig. 1. Deep Learning Architecture of AlexNet.

With ⍰ the weighted neuron input. The second layer was an Overlapping Max Pooling layer which conserved the depth, while down sampling the width and height of the tensors. The stride of 2 is used as an overlapping between adjacent windows in order to reduce the error rate. The following layer in another convolutional layer, followed by another overlapping max pooling layer, then three convolutional layers, then a max pooling layer, followed by two fully connected layers, and finally a softmax classifier with 1000 class labels. Softmax layer is a normalizing layer which that takes as input a vector of N dimension, and converts it to a N probabilities vector which is proportional to the exponentials of the input elements (2), such as the values of the resulting vector are bounded in the interval $[0,1]$.

$$\sigma(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} \qquad (2)$$

With $\sigma\colon \mathbb{R}^N \rightarrow \mathbb{R}^N$is the softmax function, $i = 1, \dots, N$ correspond to the index of the element in the N dimensional vector: $x = (x_1, \dots, x_N)$. These different layers all aim at mitigating the vanishing gradient problem during back-propagation training. Indeed, convolutional layers produce feature maps based on the idea that an object could be found in different areas of the image, and pooling reduce the neighbouring features by taking feature map maximum.

Transfer Learning approach consisted basically of reusing the values of weights of the trained Deep Neural Network, while replacing the last layers with new ones which are retrained to provide a model that is better adapted to the target object and task on hand. Moreover, it required less time and calculation power. Indeed, the initialization weights are optimal and not random. Also, as the first layers in a Deep Learning Neural Network are extracting simpler features. These features and their corresponding weights remain similar provided that the model has been pre-trained on thousands of images. Indeed, Deep Learning has thousands of free parameters, which require thousands of images of training, in order to set these parameters to optimal values, which is a drawback as this Big Data model cannot be directly applied to train Small Data applications such as the one at our hands, where the quantity of annotated data is limited. Hence, the use of Transfer Learning is essential to retain the benefits of Deep Learning, while adapting to Small Data, as it needs to retrain only the last layers which correspond to higher level information that is important for the final decision.

The choice of using Transfer Learning was motivated by our dataset size that is relatively small, which is usually the case in specific satellite image and remote sensing applications, such as sand dunes detection, while retaining the predictive power of a Deep Learning model. Also, the processing power required by Deep Learning was enormous compared with Transfer Learning, which doesn't require as much iterations to converge. Thus using Transfer learning required less CPU cycles and less time.

B. Proposed Method

The proposed method started by an image pre-processing, followed by an image clustering, then Transfer Learning:

1) Satellite image pre-processing: Satellite images are usually impacted by noise factors which could be optical or systemic. The use of a noise reduction strategy was essential. The proposed approch started by using a bilateral filter, represented by equation (3), which smooths the image while preserving edges, with a nonlinear combination of neighbouring pixel values.

$$I^{filtered}(x) = \frac{1}{W_p}\sum_{x_i \in \Omega} I(x_i)\, f_r\,(\|I(x_i) - I(x)\|)\; g_s(\|x_i - x\|) \qquad (3)$$

$$with\!:\; W_p = \sum_{x_i \in \Omega} f_r\,(\|I(x_i) - I(x)\|)\; g_s(\|x_i - x\|)\,)$$

I was the original image and $I^{filtered}$ to the filtered image. x corresponded to the coordinates of the pixel. Ω was the window centered in x, $x_i \in \Omega$ was a neighboring pixel, f_r was a Gaussian function for smoothing differences in coordinates and W_p was the weight assigned with the spatial kernel g_s. Then histogram equalization was used to enhance the contrast of the image, thus evenly distributing the intensities of pixels on the histogram. These operations were applied on both multispectral and panchromatic images. Indeed, high resolution satellite images have usually two components: multispectral images providing spectral resolution with medium spatial resolution and a panchromatic image with high spatial resolution. A fusion was operated between multispectral and panchromatic images with a bilinear interpolation to increase the spatial resolution of the multispectral images using the panchromatic image.

2) Image clustering: An unsupervisied approach was used for the segmentation of sand dunes, which consisted of two steps: the first one was the combination of multispectral bands of the satellite image into an RGB image, followed by a conversion to the CIELAB color space, which was defined by the International Commission on Illumination (CIR) with L being the luminosity layer, and a and b the chromaticity-layers encoding the color information. In the second step, a Lab image was taken as input and operated a clustering based on K-Means segmentation to distinguish between different geological aspects of the area of interest, including sand dunes. Indeed, multispectral bands in satellite images allow to get information about the nature of the terrain using its reflectance in different wavelenghts. Specifically, the distinguishment was made between different mineral compositions of sand from the satellite images.

3) Transfer learning: The proposed approach started by creating a database by annotating 1000 image patches of 228x228x3 containing three classes: barchan dunes, dunes collisions, and bare field. This particular size was selected to fit our object of interest which was barchans and also was close to the input layer size of AlexNet Deep Neural Network which was 227x227x3. Furthermore, the dataset was augmented by performing random flips along the y axis, and resized the images and scaled the patches. Our dataset was then split into 70% of images for learning and 30% for testing. Finally, the last three layers of AlexNet were replaced by a fully connected layer, a softmax layer, and a classification output layer.

4) Combining clustering with transfer learning: The segmentation labeled each pixel with a class corresponding to a specific geological feature, including sand. Clusternig results were used as a binary mask representing sand, which could be part of a barchan dune, dunes collision or even sand trail left by a moving dune. Hence, the mask was used in combination with the information resulting from Transfer Learning, which classified the image patches of the terrain in three categories: Barchan dunes, dunes collisions or bare field. The information from the clustering and Transfer Learning was fusioned as following: The Transfer Learning was operated only in areas which the segmentation results classified as containing sand. Besides, the results of the clustering were validated when they were consistant with the results found by Transfer Learning. Indeed, when the clustering indicated that the pixels in the image patch contained sand, but the Transfer Learning classified it as bare field, the image patch was declared as bare field, as sand in this case corresponded to sand trails left by moving dunes.

III. Experimental Results

A. Satellite Image Dataset

The experiments were conducted on an IKONOS high resolution satellite image, which consisted of 5 bands (Table.1). The high resolution image was taken by the IKONOS satellite on July 23rd, 2003, in the South of Morocco, near the city of Tarfaya, in the Sahara Desert, with coordinates in between 27°26'8.6621"N, 13°08'5.2628"W and 27°41'1.0350"N, 13°22'0720"W.

Two data flows were used: In the first one, the entire image was used to feed the clustering based image segmentation algorithm. In the second one, 1000 patches of 228x228x3 were selected from the satellite image to be used for Transfer Learning after annotation. The dataset was split using a ratio of 70% for the training set 30% for the test set. The next section presents both quantitative and qualitative results.

B. Image Pre-Processing

This step aimed at enhancing each band image so as to improve subsequent processing. (Fig.2) showed the results after applying histogram equalization, then the bilateral filter. One can notice that the original image was very dark, with a poorly distinguishable objects of interest.

TABLE. I. Spectral and Spatial Resolution of Ikonos Image

			Spatial resolution	**Wavelength**
Bands	Panchromatic		0.82 m/pixel	0.526 - 0.929 µm
	Multi-spectral	Near Infrared	4 m/pixel	0.757 - 0.853 µm
		Red	4m/pixel	0.632 - 0.698 µm
		Green	4m/pixel	0.506 - 0.595 µm
		Blue	4m/pixel	0.445 - 0.516 µm

Fig. 2. Image Pre-Processing Results. from Left to Right: Raw Image Sample, Histogram Equalization, Bilateral Filter.

After enhacement, one could clearly observe the two dunes in the righmost image. Multispectral bands were also fusioned with the panchromatic band to improve the spatial resolution.

C. Clustering-based Image Segmentation

The proposed approach started by experimenting with different number of clusters K={2,3,4,5,6} in order to find the optimal number for our application. A set of 3 multispectral images was randomy selected, then converted to Lab color space, in order to isolate the chromatic information in the layers a and b. K-Means results can be observed in the first row of (Fig.3): For K=2 and K=3, classes were undersegmenting the image. Indeed, although, the central corridor of dunes was well segmented, however, the dunes it contained, which were considerd as foreground, were not distringuishable from their immediate surroundings, which were considered as a background in our application. For K=5 and K=6, classes were oversegmenting the satellite image as dunes, which were our object of interest, were split into several regions corresponding to different luminosities caused by the reflection of sun on their curved surfaces. Hence, the optimal number of clusters selected for the subsequent processing was K=4, as it allowed to classify image regions into 4 categories corresponding to distinguishable geological items: The first 2 classes were considered as a background for our application, as they corresponded to solid bedrock. The third and fourth classes corresponded to sand from two different corridors. The main insterest of our application was the main dune corridor which contained the most important number of barchan dunes and was in the center of image. The secondary dune corridor which had a different mineral composition thus a different spectral response, was partially visible and felt mostly out of the limits of our satellite image borders. Therefore, the class which would be used as a mask was identified, then the process continued to find the best 3-channel combination of the 4 multispectral images, which provided the best segmentation results. For that purpose, the experimentation considered all 4 possibilities of multispectral bands 3-channel combinations: GNB, GRN, RGB, and NBR which were showed in the second row in (Fig.3). The subsequent evaluation was conducted by selecting a patch sample from the satellite image representing the most problematic area in the satellite image, as it contained barchan dunes, their collisions and bare field, including a road. A ground truth image of this evaluation patch depicting dunes was also used. The results of the statistical evalutaion of the 4 satellite images combinations were shown in (Table.2) which contains a confusion matrix and the following statistical indicators: accuracy, precision (or positive predictive value), F1-score, recall (or true positive rate), specificity (or true negative rate) and Youden's J-Statistic. The indicators showed the combination which demonstrated that the best segmentation was made using the RGB combination, having a 70% accuracy, 75% recall and a J-statistic of 0.44.

D. Transfer Learning

AlexNet Deep Learning was used as a basis of the transfer learning. After annotating 1000 image patches extracted from the satellite image and labeled as barchan dunes, dunes collision or bare field, the dataset was augmented before launching the backpropagation algorithm with the training showed in (Fig.4). After 530 iterations on 10 epochs, using a minimum batch size of 10 images, and a validation frequency of 100 iterations, an accuracy of 84.01% was obtained. The training set and testing set examples were showed in (Fig.5).

TABLE. II. SEGMENTATION EVALUATION OF BANDS COMBINATIONS

Statistical Evaluation		**Multispectral Band Combinations**							
		GNB		GRN		NBR		RGB	
		Actual class		Actual class		Actual class		Actual class	
		sand	not sand	sand	not sand	sand	not sand	sand	not sand
Predicted	sand	10081	22644	8159	28208	10270	20599	9970	16527
	not sand	3157	30167	5079	24603	2968	32212	3268	36284
Total		13238	52811	13238	52811	13238	52811	13238	52811
Accuracy		0.6094		0.496		0.6432		0.7003	
Precision (PPV)		0.3081		0.2244		0.3327		0.3763	
F1-Score		0.4387		0.329		0.4657		0.5018	
Recall (TPR)		0.7615		0.6163		0.7758		0.7531	
Specificity (TNR)		0.5712		0.4659		0.6099		0.6871	
Youden J-Statistic		0.3327		0.0822		0.3857		0.4402	

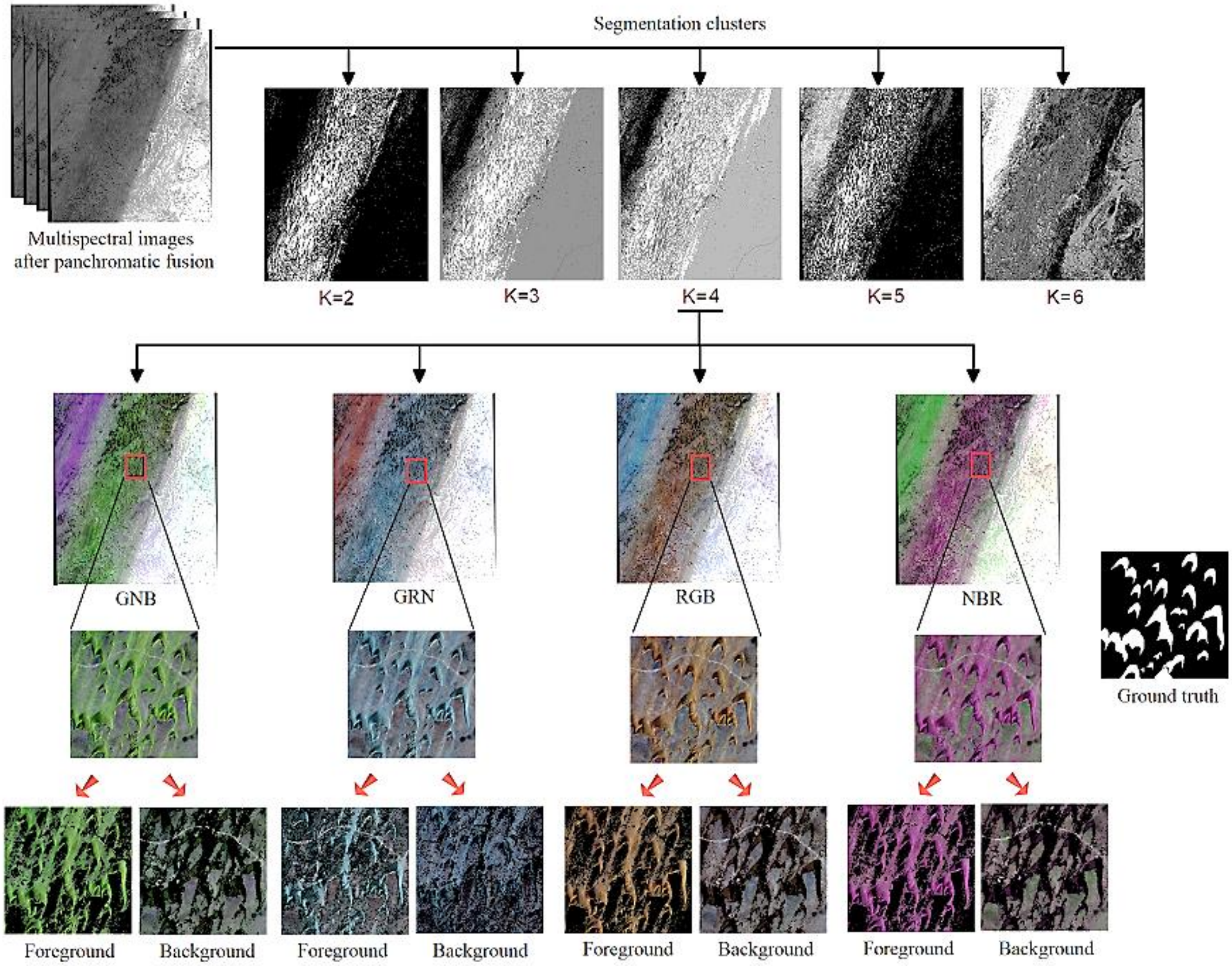


Fig. 3. Segmentation of Satellite Image Band Combinations, Evaluation Patches and Ground Truth.

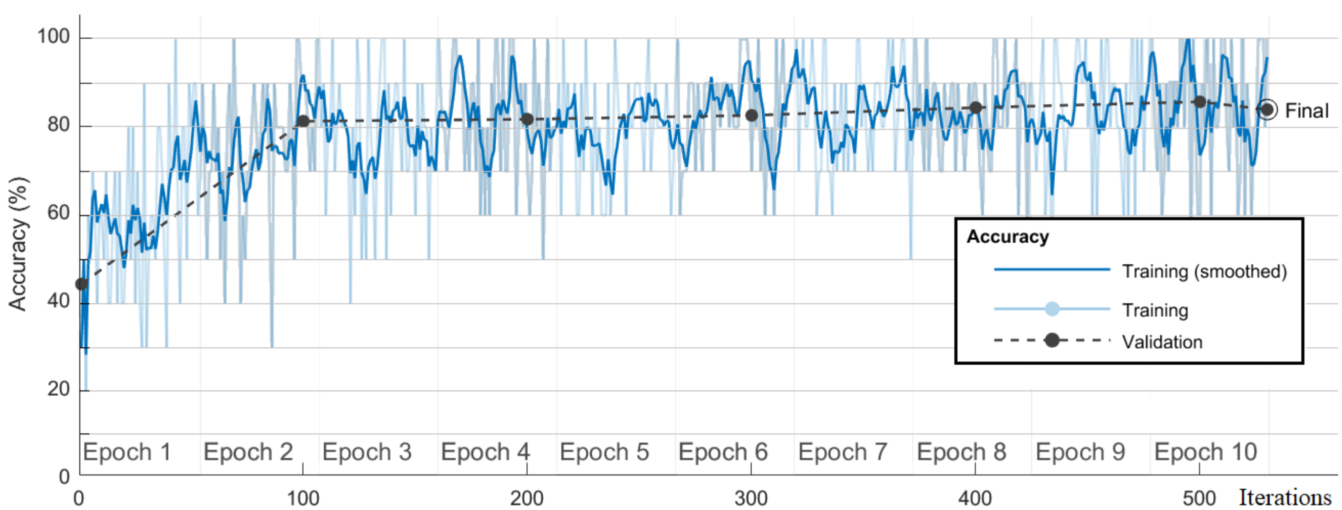


Fig. 4. Transfer Learning Training.

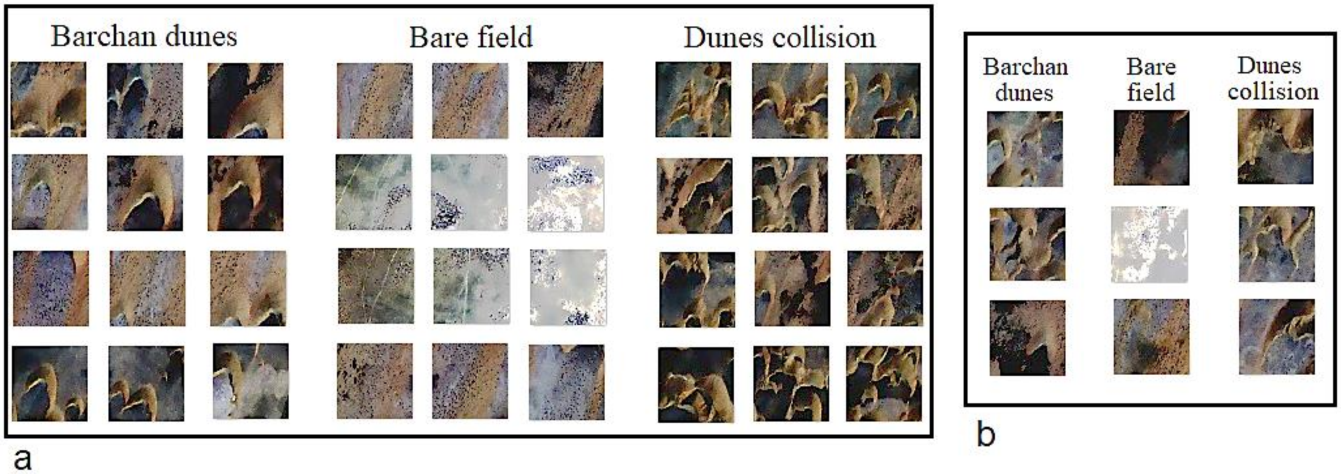


Fig. 5. Image Patches from the Satellite Image Representing 3 Classes of Objects (a): Training Set Examples, (b): Test Set Examples.

IV. Discussion

The dataset preparation benefited from data augmentation. During the first step, image enhancmened allowed data to be normalized but also enriched the multispectral bands after their fusion with the high resolution panchromatic band. On the second step, the primary segmentation eliminated the large surrounding areas containing the solid bedrock with an accuracy of virtually 100%. the secondary segmentation results obtained (70% accuracy) were also good considering the fact that they were evaluated against the patch showing the most problematic part of the satellite image, and also that the ground truth image used was approximative. Moreover, the secondary segmentation was targeting sand, while the ground truth depicted sand dunes only. Indeed, the evaluation patch at this stage, provided only with a reference to compare between the band combinations. The result was the selection of RGB combination as a basis for the third step. During this last step, one could notice that the misclassifications happened mainly between barchan dunes and dunes collisions classes as it was sometimes difficult to classify partially occluded barchan dunes. Conversely, the detections of bare field, including patches with sand trails were made seemlessly.

V. Conclusion and Future Work

In this paper, a new method was proposed for the segmentation of barchans sand dunes and their collisions in high resolution satellite images. The contribution in this paper was the introduction of a new class consisting of barchans dunes collision, and the use of Transfer Learning as a variant of Deep Learning to achieve good results. Indeed, by combining a clustering method with Transfer Learning, an accuracy of 82.01% was obtained. This information is of interest to specialists studying sand dunes dynamics, which could come up with solutions which applicable at a large scale to protect the populations and the environment. In future work, the proposed method could benefit from the use of multiple high resolution satellite images, thus providing a multi temporal analysis of sand dunes progression. Also, the proposed approach could be improved in the future by taking into account the fact that Barchan sand dunes are 3 dimensional objects, which appeared as 2D crescent like objects in satellite images. A DSM (Digital Surface Model) could be generated from high resolution satellite images showing barchans dunes, their collisions and various desert features which could help understand the dynamics of desertification.

References

[1] UNCCD: United Nations Convention to Combat Desertification in those countries experiencing serious drought and/or desertification, particularly in Africa. UNCCD explanatory leaflet, UNCCD Secretariat, Bonn 2008.

[2] Tsoar, H., 2001. Types of Aeolian Sand Dunes and Their Formation. Geomorphological Fluid Mechanics. N.J. Balmforth and A. Provenzale (Eds.): LNP 582, pp. 403–429.

[3] Elbelrhiti, H. and Hargitai H. Encyclopedia of Planetary Landforms, 2015, Chapter: Barchan, p123.

[4] M. A. Azzaoui, M. Adnani, H. El Belrhiti, I. E. Chaouki, L. Masmoudi . Segmentation of Crescent Sand Dunes in High Resolution Satellite Images using a Support Vector Machine for Allometry. International Journal of Advanced Computer Science and Applications. 2019. ISSN:2158-107X. Pages 191-198 Vol. 10, No. 11.

[5] McKee, E.D. Introduction to a study of global sand seas. McKee, E.D. (Ed.), A Study of Global Sand Seas: United StatesGeological Survey, 1979. 1052, pp. 3–19.

[6] Wolfe, S.A., Hugenholtz, C.H. Barchan dunes stabilized under recent climate warming on the northern Great Plains. 2009. Geology 37, 1039–1042.

[7] Kocurek, G. A., & Ewing, R. C. Aeolian dune field self-organization - Implications for the formation of simple versus complex dune-field patterns. Geomorphology. 2005. 72(1-4), 94-105. https://doi.org/10.1016/j.geomorph.2005.05.005

[8] Hugenholtz, C. H., and T. E. Barchyn. Spatial analysis of sand dunes with a new global topographic dataset: New approaches and opportunities, Earth Surf. Processes Landforms. 2010. 35, 986–992, doi:10.1002/esp.2013.

[9] Blumberg, D.G.. Remote sensing of desert dune forms b. polarimetric Synthetic Aperture Radar (SAR). Remote Sensing. of Environment 1998. 65, 204–216.

[10] Vermeesch, P., Drake, N., Remotely sensed dune celerity and sand flux measurements of the world's fastest barchans (Bodélé, Chad). Geophysical Research Letters. 2008. 35, L24404.

[11] Bishop, M.A.,. Comparative nearest neighbor analysis of megabarchanoid dunes, Ar Rub al Khali sand sea: the application of geographical indices to the understanding of dune field selforganization, maturity and environmental change. Geomorphology 2010. 120, 186–194.

[12] T. Blaschke, Object based image analysis for remote sensing, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 65, Issue 1, 2010, Pages 2-16, ISSN 0924-2716, DOI: 10.1016/j.isprsjprs. 2009.06.004.

[13] Schmidhuber, J. Deep learning in neural networks: an overview. Neural Networks 2015. 61, 85–117.

[14] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Sanchez, C.I. A survey on deep learning in medical image analysis. Med. Image Anal. 2017. 42, 60–88.

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. Generative Adversarial Networks. 2014. arXiv:1406.2661.

[16] Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014. Available: https://arxiv.org/pdf/1409.1556.pdf.

[17] He, K., Zhang, X., Ren, S., Sun, J.. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, pp. 770–778.

[18] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17). 2017, pp. 4278–4284.

[19] Chen, Y., Lin, Z., Zhao, X., Wang, G., Gu, Y. Deep learning-based classification of hyperspectral data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2014. 7 (6), 2094–2107.

[20] Zhu, X.X., Tuia, D., Mou, L., Xia, G., Zhang, L., Xu, F., Fraundorfer, F. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 2017. 5 (4), 8–36.

[21] Zhang, L., Zhang, L., Du, B. Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Geosci. Remote Sens. Mag. 2016. 4 (2), 22–40.

[22] Krizhevsky, A., Sutskever, I., Hinton, G. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. 2012. pp. 1097–1105.

Usefulness of Mobile Assisted Language Learning in Primary Education

Kashif Ishaq*[,1], Nor Azan Mat Zin[2], Fadhilah Rosdi[3]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Malaysia

Adnan Abid[4], Qasim Ali[5]
School of Systems and Technology (SST)
University of Management and Technology
Lahore, Pakistan

***Abstract*—Literacy & Numeracy Drive (LND) is a mobile application that is used in public sector primary schools in Punjab province, Pakistan to teach students of Grade 03 on a tablet for learning languages and Mathematics. Persons designated the role of a Monitoring & Evaluation Assistant (MEA) visit every school allocated by authorities once in a month and select 07-10 students randomly to evaluate them on his own tablet by asking multiple questions related to English, Urdu and Mathematics. After the evaluation, MEA has to upload the result on the official portal for the respective school. This study aims to evaluate the effectiveness of LND for its usefulness, usability, accessibility, content, and assessments by involving students and teachers using this application in different schools. A mixed-method study has been adopted in which 57 teachers and nearly 300 students from different locations of the district and from different schools have been selected, to measure the effectiveness of LND and evaluate the effectiveness with the help of interviews and questionnaires. The result reveals, in its current form, the LND application is not effective and needs improvement in usability, design, content, accessibility, infrastructure, and assessment. Furthermore, teachers recommend that game-based learning consists of an interactive interface, phonics, animations. As the more interactive and attractive presentation of the content and variations in the assessment may increase students' involvement and will make this application more effective and will produce good results.**

Keywords*—Literacy and numeracy drive; monitoring and evaluation assistant; assessment; usability; content; design; infrastructure

I. Introduction

"English is not a subject which can be taught; it is a subject which must be learned (Michel West)." English is a world language and is being schooled round the world. It is an efficient method of communication. Native speakers learn it as a primary language and several nations taught this as a second language [7]. English is the language of instruction in schools at all levels [25][2]. Most states have realized the significance of the English language in teaching. Applying modern tools to support students learning English is a critical issue where English is not adopted as a primary language [18]. An important requirement of the day is to learn English as a second language to survive in the global community [8]. It is also observed that having a strong English background supports the candidates for entry into the higher education institutions and assists them to have better job opportunities [16][46].

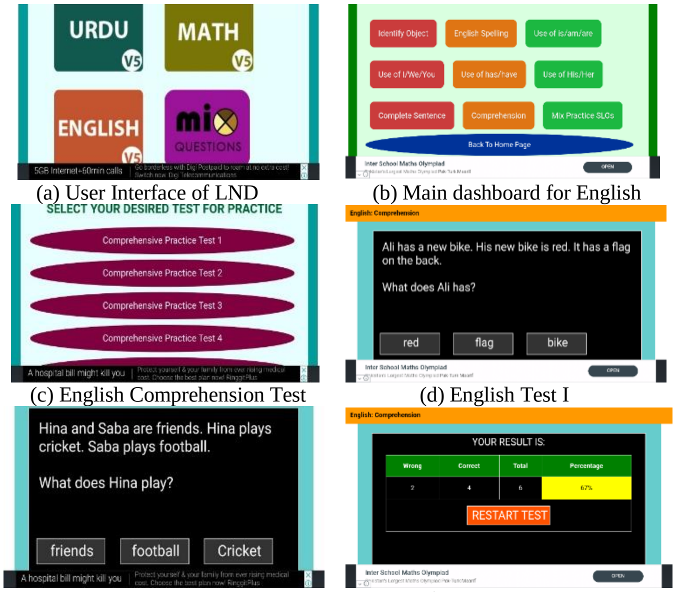
A. Preliminary Section

1) Literacy and Numeracy Drive (LND): Outdated procedures of large scale assessment methods for primary students (such as Punjab Education Commission Examinations) are expensive, multifaceted, and irregular. In 2015, the Department of School Education and the Punjab Information Technology Board (PITB) implemented mobile-based application which is low cost named as Literacy and Numeracy Drive (LND) for teaching English, Urdu and Mathematics in all public sector schools of Punjab Province, Pakistan as well as used by Monitoring Officers of School in their scheduled visits of every month in every public school. The application is connected to a widespread question bank and each question is marked with the appropriate student learning results [39]. The main user interface is shown in Figure 1.

To achieve better learning outcomes as it has been mentioned earlier that School Education Department, Punjab has employed LND Application for assessment of learning outcomes of grade three students on a monthly basis. For this purpose, students are educated according to the Prescribed Students Learning Outcomes (SLOs) of three basic subjects on a mobile-based application i.e. English, Urdu and Mathematics [39]. But the major focus of this research is only on the English subject because the English language is a challenge for teachers to teach to the students at all educational levels because it is not their native language and used as a secondary language [60]. Schools in rural areas in Pakistan have no/less access to technology through which they may get help teaching English and they keep on applying conventional teaching methods which only develop cramming and rote learning among students and they remain passive and reluctant to learn English.

2) The Assessment Process of LND: The Monitoring and Evaluation Assistants (MEAs) have been trained to select 7-10 students of grade 3 randomly and to give them the spot test on their own tablet. For the particular grade, theme and key student learning outcome (SLO), MEA ask random questions from established question bank from the randomly selected students to complete the assessment process [78].

*Corresponding Author



(a) User Interface of LND (b) Main dashboard for English
(c) English Comprehension Test (d) English Test I
(e) English Test II (f) LND Test Assessment

Fig. 1. LND Interfaces.

Across the province of Punjab, there are more than 52000 public sector schools in which this assessment is conducted on a monthly basis by asking 07 questions in total from every student. This process takes less than 05 minutes for each student and approximately 329,000 students are evaluated monthly. According to [53], 6.7 million evaluations have been conducted till now by the MEAs and this data is uploaded on the official portal for the analysis to the educational administrators via SMS and also online dashboard whereas, the monthly assessment result is available for free for every school in every sub-division of all the districts in Punjab province for their information [78][80].

II. Literature Review and Theoretical Framework

The English language functions like blood in the veins of states worldwide. As a result, the need for the English language has expanded to the level that it involves that citizens of modern societies to be adept enough in order to be successful in their academic and professional careers [20]. A survey conducted by [19] in 89 countries over the globe to identify the qualification and training method for English language teachers to teach English, in which the results revealed that teachers are not specially trained to teach English and it's not the problem of one country or a region but this is the problem in all over the world. The ever-growing need for the English language in different spheres of human life, that's education, financial matters, legislative issues, excitement, innovation, and trade, has made a huge appeal for English language teaching around the world [20].

A. *Technology Acceptance*

Davis' (1989) TAM (Technology Acceptance Model) is a well-known model of research in technology admissibility over the past two decades [41][66]. The TAM is developed to describe the predictors of a range of computer knowledge through different people [70], based on a (Fishbein and Ajzen's 1975) principle of rational comportment. The conventional element of TAM [68] is five: perceived utility, ease of use, usage attitudes, shared intention and discrete use. Several researchers have authenticated TAM [50] and TAM can account for approximately 40% of the device use, according to [69]. Nonetheless, there are a number of aspects that the template has played a crucial role in forecasting its use [34].

That is why the single TAM [81] was updated and the Unification Theory for Technological Entry and Use [51] were also transferred to TAM2 [71]. Extended TAM variants to test the adoption of consumers of a range of learning skills and in various settings including information management systems and smartphone leaning were developed [1].

B. *Perceived Ease of use*

The easy use of computer programs in the TAM is an essential determinant of the goal. A simple use is defined by a discreet person as being simple in implementing a process [82] and as having a direct positive effect on the perceived usefulness. Nonetheless, the traditional TAM [22] implicitly affect the purpose of using a method by using behaviors against the use of certain experiments [50][69]. In addition, several studies have not established direct links between user-friendliness and user motivation [47][63][52] claiming to be involved with the work of technology amongst these parameters [63]. [9] suggests that it is hard for the systems to produce any degree of satisfactory disappointment in such places to allow the results of mastering and game technology simple to use [6]. The results of the research [24] can lead to the conclusion that supposed ease of use does not lead to the gratification of digital game students. Nevertheless, [9] found that game preferences are easy to use.

C. *Perceived usefulness*

The use of term performance throughout this interpretation may, however, indicate a focus on the result/result during the application process. In [82] is another important indicator of readiness to make use of the program within TAM. The supposedly "to the stage that one supposes that use of certain devices will enhance their educational outcomes." It may be beneficial in many respects, but the learning process itself might be as important or valuable as the learning outcome. [74] presented external opportunities for growth.

D. *Technology in Education*

Technology is a gift of God and the mother of community formations, expressions, and sciences. It has affected distinctive features of life and re-imagined living. The subject of education has been revolutionized by technology and its importance in schools is being considered. It has worked out to be simpler for teachers to get the learning of personal computers (PC) in teaching. The business of technology in the area of education is four-overlap: it is incorporated as a part of the educational programs, and instructional delivery framework, a method for holding instructions and an apparatus to upgrade the whole learning process [55]. Recently, with the growth of new modernized and net-based advancements students may interface with the web and sign in to explicit sorts of net recreations to take an interest effectively in interactive games [5].

Research into the utilization of technology by students is an abandoned area in Pakistan. Most of the researches in Pakistan investigated the utilization of social networking websites by students and educators in teaching and non-teaching settings and the general utilization of the web by tertiary students [57]. The utilization of technology in students' lives has expanded so much that it is needed to show students the method of teaching sentence structure through visuals and motion pictures [12]. Teachers consider the appropriation and usage of Information & Communication Technology (ICT) in teaching since they are the key to making learning take place [61][30].

E. Teaching English as a Second Language (TESL) in Primary

In Pakistan, the English language is taught as a second language. Our instructional and official languages are English and Urdu. In Pakistan English is fast using language there are 11% Pakistani who speak English which makes Pakistan Asian’s 3rd largest country [7]. It has its challenges especially in rural area schools because they feel difficulty understanding it because they are bound to learn it as a second language. Rules of English grammar are very difficult for them to understand. They belong to the socio-cultural background where English is not spoken as a primary language. They speak Punjabi at home and Urdu at school whereas English is taught as a secondary language. It is also reported here that difficulties in teaching English are very much related to students’ learning [7]. Many factors, parents’ socio-cultural background, teacher professional development, curriculum, school infrastructure and facilities contribute to the achievement level of students in English language learning. The only change in the medium of instruction does not make a difference. All related factors need to be addressed in terms of improving English language learning [14].

Globally countries around the world are boosting teachers to use digital resources in language learning because in traditional classrooms teachers and students only communicate in the classroom whereas the learning environment has been changed with the emergence of iPad, iPhone or other Android applications [73]. According to [42] digital technology enables the students to communicate in an environment to enhance motivation in learning a subject. A study conducted by [27] on the use of learning products in the global market for digital English language. They mentioned that its use reached up to \$1.8 billion in 2013 and it is proposed that revenues would reach over \$3.1 billion by 2018. Use of Apps like Duo Lingo is enormously widespread for language learning, with over 70 million sign-ups [17].

F. Mobile-Assisted Language Learning (MALL)

[67] mentioned that smartphones, mobiles, e-reader, tablets, handheld gaming consoles, portable audio players, laptops, Netbooks and console are effective mobile learning and teaching tools. Mobile Applications for Language Learning (MALL) are delineated as utilization of mobile phones in language teaching and language learning [3][40][79]. The mobile phone enables students to learn casually to improve their learning abilities and language reading [54]. Like western countries, mobile devices are extensively being used in developing countries as well for educational purposes [38][76]. Mobile technology has brought a very positive change in pedagogical approaches. Incorporating mobile devices with the curriculum enables the learners to learn things on an independent level free of time and space by applying the latest software like simulations, learning games, online assessment, language learning applications [31]. A smartphone is used for various kinds of activities like watching movies, social networking, gaming, transfer of remittance, accessing desired information, online admissions and many more [37][3]. 1.2 million Individuals in 2012, exploit mobile applications everywhere throughout the Sphere and in 2017 it is predictable that the amount will be 4.4 million [62]. Mobile learning is a service that gives universal information automatically to the learner and makes the learning content available which helps the success of knowledge [28].

As compared to traditional computer devices, learners are enabled to take the advantages of segmented time to learn without any restriction of time and location with the help of mobile technologies’ affordability, movability and user-friendliness [76] because low cost, flexibility, small size, user-friendliness and allowing more sophisticated are the advantages of MALL which are rapidly attracting new users [29]. Likewise, mobile phones can offer language students with understandable info by means of pre-customized programming, offer access to chances of exchange of significance through connecting with the software or instructors, and build a student-focused learning environment [76].

Past research demonstrates that MALL influences vocabulary learning systems, self-governance, and reading comprehension [26]. Similarly, to access new content-on-demand there is a need for Internet access that allows the user to do so whereas it requires the capability for doing such is pointed by [43]. Informal learning may contribute to the habit of personal education and environmental possibilities, including social and technological influences, taking account of the student [72].

On the other side, there may be a question that digital education is not ready to try modern new technology [13]. In App-Store (iOS), Play Store (Android) and other repositories, mobile applications are available. They are rising continuously. Such apps also include strengths and weaknesses in terms of application characteristics, offline functionality, discoverability, speed, deployment, maintenance, and autonomy of networks, limitations to content, approvals method, charges, development costs and user interfaces [11].

Another task involving professional skills and contours is the development of m-learning apps. The progress of learners is either defined by the way that the environment is translated into software or by the system change (not innovation) for the utility of m-learning resources [62]. [72] says that when implementing mobile training, a program programmer is required to follow 10 basic guidelines. Sadly, as a whole, electronic literacy model is an obstacle for both teachers and learners [10], as technological discrimination and behaviors are another issues, such as adaptability [65], usability [44], and subscription fee [44], etc.

G. Challenges of English Language Learning in Pakistan

The education system in Pakistani can be categorized according to the medium of education which corresponds with socioeconomic classes. The education system comprises of government, private, and religious schools (Islamic seminaries named madrasas) where Urdu is most commonly used as intermediate of instruction whereas, in private and elite institutes, English is the medium of instruction [32]. The prime challenge for the English language is multilingual classrooms, facing by the teachers as well as students not only in Quetta, Pakistan but also in other parts of the globe. Due to the different local languages spoken by the students in the classroom, teachers and students have to face a number of problems because it becomes a big challenge for English teachers to deal with students speaking different languages in the classrooms [56].

[33] researched the many challenges facing English language teachers and students in the multilingual schools in District Mianwali and District Bhakkar, Punjab Pakistan. The respondents were teachers and students who revealed, due to the incompetent instructors, educational policy and the test process and lengthy curricula, that college students find it hard to learn English to speak in. The need for language skills in the context of the English curriculum is not recognized at any stage of education in Pakistan, as when students reach the university level they have to face problems with English topics, even if they are adopted as a mandatory topic at the school level. With regard to schools in the public sector, the traditional method for the teaching of English is employed and the government's recommended curriculum is not used to test or teach speaking skills [48]. Language is called an entity, the primary function of which is to facilitate human communication. The development and implementation of the ESP (English Syllabus for Pakistan) mandated that English be used to educate teachers and use communicative/interactive methods [48]. The position of English in Pakistan is especially complex because the government's language, military and higher education are, and even the language of control and the language of the ruling class before independence ruled the country [14].

Some writing skill techniques, such as, brainstorming, outlining and pre-writing discussions on the topic is familiar by some English language teachers but numerous causes that comprise lack of time and learners' weak educational background, restricted curriculum, teacher's lack of command on reading skills and lack of training are considered challenges for teaching reading skills, due to the most of the teacher do not apply these approaches in classrooms. Some of the major encounters faced by instructors in teaching reading skills, students' lack of vocabulary, interest in reading activities and a large classroom [20]. The medium of instruction is English and Urdu in schools but teachers mostly use Gujari and Shina for primary grades which were remarked by the representative of Gujari. In some areas, Gujari is used as a support language for elementary and matric grades as well [14].

There is very poor learning quality throughout the country, especially in public and low-cost private schools. The assessment surveys of students at the national level have shown continued low mastery in science, math, and languages throughout the years. Lower learning results are tied directly to teachers ' low education quality, as the main reasons for teachers in Pakistan are their low level of familiarity and poor pedagogical knowledge. According to [49] Pakistani schools are facing the deficiency of highly qualified subject expert teachers especially in English, Science, and Mathematics through all tertiaries of Pakistan but English as medium of instruction used by the private schools predominantly because children in this sector have strong focus on the use of English language, while students in schools of public sector mostly their regional mother tongue or Urdu as a medium of education. Although, many public sector schools have the medium of instruction in English children are incompetent to accomplish even basic proficiency levels of the language because of the low aptitude of instructors for teaching English as a second language, [49].

By finalizing the literature related to MALL based interactive applications, it is concluded that there is no research has been conducted for the effectiveness of LND in public sector schools for grade 3 student of District Sheikhupura, Punjab, Pakistan and this will be very 1st study on this mobile-based application which is currently adopted by school education department for teaching English, Urdu and Mathematics.

Research Objective:

- To identify the factors that affect usage of MALL based tool (LND) for TESL in Pakistan.

Research Question:

- To what extent the teachers in public sector schools have the accessibility of the Internet and smartphones?
- What are teachers' perceptions of LND for its usefulness, content, usability, accessibility, and assessments?
- What are the potential advantages and disadvantages of LND from the students' perspective?

III. Methodology

A. Setting and Sample

This study took place in public sector primary schools of Pakistan which is consisted of four provinces, Punjab, Khyber Pakhtunkhwa, Balochistan, and Sindh. The Punjab province is the most populated area of Pakistan selected for this study which is further divided into 36 districts. It has 52,394 schools (male and female), 12,268,981 students (male and female) and 403,172 teachers (male and female). Furthermore, District Sheikhupura is selected for the survey and interviews of teachers and students which is sub-divided into 05 regions (Ferozwala, Muridke, Safdarabad, Sharaqpur, and Sheikhupura). It has 1,247 schools (male and female) out of which 21 schools were randomly selected to visit in all regions. This study is to measure the effectiveness of LND for English subjects only so, all the teachers in selected schools are the subject expert of English, who are using LND mobile application in class for teaching.The sample was drawn from teachers of 21 randomly selected schools out of 1,247 from District Sheikhupura and 57 teachers who were teaching English subjects and also using LND application in class and 300 students of grade 03 only were invited to participate in the study by accomplishing questionnaire and interview in class timing. The rate of response was 100% because all teachers participated happily.

B. Participants

For the study, 57 respondents participated from which male respondents were 12 (21.1%) and the majority of respondents were female which was 45 (78.9%). The age group of 18 (31.6%) participants were 26-30 which is the majority whereas 09 (15.8%) respondents having the age group of 31-35 whilst 13 (22.8%) and 17 (29.8%) respondents have the age group of 36-40 and above 40 respectively. From the students' side, 300 participants were selected in which 138 were male which was 46% and the rest were female with 162, which corresponds to 54%, out of a total of 100%.

C. Instrument

The questionnaire for teachers includes a total of 72 questions and reliability was measured by Cronbach's alpha which was (.805) that indicates a high level of internal consistency. The questionnaire includes nine factors: demographic information, ICT literacy, Design, Usability, Accessibility, Usefulness, Content, Assessment and Provider of application. The questionnaire for students includes a total of 30 questions including demographic information and LND questions and reliability was measured by Cronbach's alpha which was (.865) that indicates a high level of internal consistency.

In the section of demographic information: age, gender, education, and location of the school was asked from the participants whereas in ICT Literacy: access of a computer, mobile devices/tablet and its usage that had yes, no options were asked. In the section of Design of LND: icon of application, attractiveness of screen and buttons, easy to use interface, font size, navigation keys, voice instructions, animations, videos, recommendations of questions, question bank and advertisements; in the section of Usability of LND: easy to find icon, easy to use interface, easy to use touch screen, easy to use input, step by step assistance, difficulty, performance and improvement; in the section of Accessibility: equal access to use the application and equal time for practice; in the section of Usefulness: enhance of vocabulary, saves my time, pictorial presentation, improvement of knowledge, skillful in learning and confident in speaking; in the section of Content: relevancy of content, suitability of content with curriculum, development of interest, suggestion for content and improvement in content; in the section of Assessment: communication of results to parents, suggestions for answers, recommendations of questions on the basis of answers, improving learning outcomes, weakness of the students, and fairness of assessment; in the section of Provider of application: feedback for application, recommendations for application and enforcement of new technology was asked where all items based on 5-point Likert scale, ranging from 1-Strongly agree to 5-Strongly disagree.

Lastly, two open-ended questions were asked about missing features in LND and what & how features should be added in LND to make it interactive, interesting and more useful for students.

D. Data Collection and Analysis

The data was gathered during the winter session in 2018 and the questionnaire was administered to all English subject teachers and students of grade 03 and the participation was voluntary and anonymous. For the analysis of questionnaire data, Statistical Package for Social Science (SPSS 25.0) was used in which frequencies were generated for demographic information of the participants whereas descriptive statistics were calculated for the ICT Literacy section. For the sections of LND, Mean and Standard Deviation were calculated accordingly. Response to open-ended questions was analyzed through open coding where first, categories were formed and then the assignment of titles was made.

IV. RESULTS AND DISCUSSIONS

A. Teachers' Result

The analysis of the questionnaire in tabular form is given in which higher the mean value shows the strongly disagree and lower the mean value shows the strongly agree.

According to Table 1. majority 18 (31.6%) respondents having the age group of 26-30 whereas 09 (15.8%) respondents having the age group of 31-35 whilst 13 (22.8%) and 17 (29.8%) respondents have the age group of 36-40 and above 40 respectively and According to Table 2. Out of 57 respondents, male respondents were 12 which is 21.1% and the majority of respondents were female which was 45 (78.9%) of the total of 100%.

According to Table 3. The majority of the respondents 40 (70.2%) have their computers at home to use but 17 respondents (29.8%) do not have computers to use at their homes for educational or entertainment purposes and according to Table 4. The majority of the respondents 49 (86.0%) have their smartphones to use whereas 08 respondents (14.0%) do not have smartphones to use in a routine matter for educational or entertainment purposes.

According to Table 5. Which describes that overwhelming majority 53 (93.0%) have Internet access but only 4 (07%) respondents do not have Internet access to use for educational purpose.

TABLE. I. AGE

	Frequency	Percent
26-30	18	31.6
31-35	9	15.8
36-40	13	22.8
Above 40	17	29.8
Total	57	100.0

TABLE. II. GENDER

	Frequency	Percent
Male	12	21.1
Female	45	78.9
Total	57	100.0

TABLE. III. DO YOU HAVE A COMPUTER TO USE?

	Frequency	Percent
Yes	40	70.2
No	17	29.8
Total	57	100.0

TABLE. IV. DO YOU HAVE A SMARTPHONE?

	Frequency	Percent
Yes	49	86.0
No	8	14.0
Total	57	100.0

TABLE. V. DO YOU HAVE INTERNET ACCESS?

	Frequency	Percent
Yes	53	93.0
No	4	7.0
Total	57	100.0

1) Design of LND: Participants disagreed with the items related to the design of LND. The majority disagree and strongly disagree with all the statements on this subscale except the item 10. The majority (98.2%) of the participants disagree with the color scheme used in the LND application and the font size is also not readable by (93.0%) that create problems while reading the text in LND application. Most importantly, all the participants strongly disagree with ($\overline{X} = 1.00\ and\ SD = .00$) that the LND application provides voice instructions/animations/videos for learning to the students. These elements can create motivation for the usage of the application. Item 10 received the highest mean score on this subscale (Table 6) in which ($\overline{X} = 4.91\ and\ SD = .29$) it clearly states that the LND application shows advertisements while using it in the classroom for teaching or at home while practicing.

2) Usability of LND: The majority of the respondents with item 13 (56.1%) disagree and (43.9%) strongly disagree, item 17 (45.6%) disagree and (54.4%) strongly disagree and item 20 (98.2%) strongly disagree where these items have a lowest mean score. These items belong to the easy to use touch screen input, assistance in difficulty and improvement of the application. However, item 18 (79.0%) agree and item 19 (75.4%) strongly agree and (24.6%) agree where these items have a highest mean score that explains application difficulty to use and performance of the application is slow (Table 7).

TABLE. VI. MEAN AND STANDARD DEVIATION FOR DESIGN OF LND

Sr. #	**Items**	**N**	**M**	**SD**
1	The icons and buttons are attractive and recognizable.	57	1.78	.57
2	The color scheme of buttons is attractive.	57	1.95	.23
3	The color scheme of the application screen is attractive.	57	1.56	.50
4	The font size is easy to read.	57	1.95	.44
5	The application provides navigation keys.	57	1.14	.35
6	The application provides useful voice instructions.	57	1.00	.00
7	The application provides animations for learning.	57	1.00	.00
8	The application provides videos for learning.	57	1.00	.00
9	The application instructs to fix the problem automatically.	57	1.16	.37
10	The application shows too many advertisements.	57	4.91	.29
11	The application provides a variety of questions in its question bank.	57	1.19	.40
12	The application provide self-recommendations for questions.	57	1.01	.13

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

TABLE. VII. MEAN AND STANDARD DEVIATION FOR USABILITY OF LND

Sr. #	**Items**	**N**	**M**	**SD**
13	The application icon is easy to find.	57	2.44	1.09
14	The application interface is easy to use.	57	2.25	.61
15	The application provides easy to use touch screen input.	57	1.56	.50
16	The application provides step by step assistance to use it.	57	2.14	.72
17	The application provides assistance in difficulty.	57	1.46	.50
18	The application is difficult to use.	57	3.77	.54
19	The performance of application is slow.	57	4.75	.43
20	The application provider is taking steps to improve the application.	57	1.02	.13

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

Respondents are unhappy with the use of LND application because of the various issues like touch screen input, having no step by step assistance to use it. It also does not provide assistance in difficulty as well as it is also difficult to use it. Most important is the performance of application which creates interest for the user to use it smoothly but in item 19, the performance of the application is slow as mentioned ($\overline{X} = 4.75\ and\ SD = .43$) that is the problem for the users while using in the class. Lastly, the application provider is not taking any steps to improve the application that affects the smooth usage of the application.

3) Accessibility of LND: According to Table 8 for Literacy and Numeracy Drive (LND) accessibility, item 21 depicts that the overwhelming majority (87.7%) where ($\overline{X} = 1.61$ and $SD = .82$) disagree that each student gets equal access to the application in class. Majority of respondents which is (96.5%) where ($\overline{X} = 1.63$ and $SD = .70$) disagree of item 22 that each student gets equal time for the practice of the application in the class because every school is allowed to have only 01 tablets in the huge strength of the class. For the usage of the LND application, it requires proper attention and time to do practice as well as to solve the exam but it is not possible in the situation explained above in the table.

4) Usefulness of LND: In the table 9, approximately all the items showing the lowest mean for the usefulness of the LND application. The application must have some proper output/outcome that could enhance knowledge and vocabulary but the above table shows (96.5%) respondents disagreed for enhancing vocabulary. It also depicts in item 26 that (96.5%) majority disagreed that LND application improves there knowledge neither it provides the pictorial presentation of concepts by (93.0%) majority in item 25 of table 9, so the respondents can make an understanding with the text easily and smoothly.

5) Content of LND: According to Table 10 for Literacy and Numeracy Drive (LND) content, the majority of the respondents disagree for the content used in the LND application. This is also depicted in the items given above with the lowest mean score. In item 30, (91.2%) majority disagreed that the content used in the LND application is taken from the curriculum defined by the Government officials neither it is suitable for learning comprehension of English, disagreed by (95.0%) majority of item 31. If the content is not suitable for learning, it creates the problems and also the cause of lack of interest for the learner to learn smoothly but item 32 indication, (93.0%) majority does not interest to learn comprehension using LND application. The last item depicts by an overwhelming majority (98.2%) where ($\overline{X} = 1.01$ and $SD = .13$) strongly disagreed that the application provider is improving the content problem with the consent of the teacher. If the provider of application uses the content with the consent of teachers and according to the curriculum then it will have more effect on teachers and learners as well.

6) Assessments in LND: The majority of the respondents disagree with the items given in Table 11, in which all the items have the lowest mean for the assessment module. If the application provides the content for learning on the basis of the result of a student then it will be very effective as respondent gets questions or topics to learn for further assessment, but item 37 shows that (86.0%) majority disagrees that LND application provides topics or questions according to the result of a student. It is also a major element, if an application itself provides the weakness by assessing a respondent automatically according to the given answers then it will create much interest and motivation towards the effective use of it. The majority of the respondents (86.0%) said the LND application is not providing the weakness of the students in item 39.

TABLE. VIII. MEAN AND STANDARD DEVIATION FOR ACCESSIBILITY OF LND

Sr. #	Items	N	M	SD
21	Each student gets equal access to the application in class.	57	1.61	.82
22	Each student gets equal time for practice of the application in class.	57	1.63	.70

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

TABLE. IX. MEAN AND STANDARD DEVIATION FOR USEFULNESS OF LND

Sr. #	Items	N	Mean	SD
23	It helps me to enhance my vocabulary.	57	1.54	.57
24	It saves my time when I use it for teaching.	57	1.47	.50
25	The application provides pictorial presentation of concepts.	57	1.07	.26
26	The application improves my knowledge.	57	1.75	.51
27	The application makes me skillful in learning English.	57	1.84	.65
28	The use of application makes me confident in speaking English?	57	1.21	.41

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

TABLE. X. MEAN AND STANDARD DEVIATION FOR CONTENT OF LND

Sr. #	Items	N	Mean	SD
29	The content in application is relevant with the course curriculum.	57	2.11	.45
30	The content in application is extracted from the curriculum.	57	1.91	.58
31	The content in application is suitable for learning comprehension.	57	1.07	.26
32	The content in application develops the interest to learn comprehension.	57	1.51	.50
33	The application suggests the content for learning comprehension.	57	1.16	.37
34	The application provider is improving content problem with the consent of teacher.	57	1.01	.13

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

TABLE. XI. MEAN AND STANDARD DEVIATION FOR ASSESSMENT OF LND

Sr. #	Items	N	Mean	SD
35	Results in application are communicated to parents.	57	1.89	.31
36	The application provides suggestions for answers.	57	1.56	.68
37	The application provides the recommendations of questions on the basis of result.	57	1.14	.35
38	Results in application helps in improving learning outcomes.	57	1.28	.49
39	The application provides the weakness of students.	57	1.25	.43
40	Assessment in application is fair.	57	1.88	1.09

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

TABLE. XII. MEAN AND STANDARD DEVIATION FOR PROVIDER OF LND APPLICATION

Sr. #	Items	N	Mean	SD
41	The application provider asks you about the feedback of application.	57	1.00	.00
42	The application provider asks you about the recommendation regarding application.	57	1.00	.00
43	The application provider enforces new technology of teaching in schools without consulting?	57	5.00	.00

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

7) Support from the Provider: Feedback is the major element that helps to improve the quality of the application or product by getting it from different people. In Table 12, the majority (100%) of the respondents strongly disagreed that the provider of LND application asks you about the feedback where ($\overline{X} = 1.00\ and\ SD = .00$). The recommendations also have the effect as feedback but in LND, all the respondents (100%) where ($\overline{X} = 1.00\ and\ SD = .00$) strongly disagreed that the application provider asks the respondents for the recommendations about the LND application improvement. Lastly, all the respondents (100%) where ($\overline{X} = 5.00\ and\ SD = .00$) strongly agrees that the application provider enforces new technology of teaching in schools without the consulting the respondents that have an adverse effect on them.

B. Teachers' Interview

A face to face interview was conducted in public sector schools in district Sheikhupura, Punjab, Pakistan only with the teachers having specialization in English subject and also have the experience of using mobile and internet. To identify the features/facilities required by teachers (best available) to teach English comprehension effectively, open-ended questions were asked from 45 (78.9%) female teachers & 12 (21.1%) male teachers to improve current LND software used in the school education department in Punjab, Pakistan.

- Animations, Voice Pronunciation and Learning Videos: Out of 45 female respondents, 39 and Out of 12 male respondents, 09 quoted to teach English comprehension will be very much effective with the help of animations, pronunciations of sounds and learning videos because if the students will be given such platform to learn English comprehension with interactive animations and voice pronunciation then they will produce more positive results.
- Design (Colorful Pictures, Icons, and Screens): According to the design perspective, 31 female and 11 male respondents give arguments that if icons and screens will be colorful and well designed, having colorful pictures in their background or pictorial form of presentation of concept will be more effective to develop the interest of students learn more.
- Content (Relevant with Translation and Meaning): For the content perspective, 21 female and 12 male respondents talked about the content of the current LND application that is not relevant to the curriculum and teachers to have to put extra effort to cover the syllabus from the book as well as doing practice on the tablet for LND software. They also said, if the content in the application will be relevant to the curriculum (also have translation and words meaning) then it will more effective for the teacher as well for students to save their time by teaching and learning from a single platform.
- Game-Based (Fun learning with Reading and Spellings): Out of 45 respondents, 21 female and 11 male respondents talked about game-based learning (fun learning) that if we teach students in the environment where they can learn by playing games then it will be much interesting for them to put more attention on learning by playing game as well as they will perform well in reading and spellings.
- Assessment (Focus for outcomes on result basis and communicated to parents): For the assessments, 12 female and 12 male respondents give their views about assessment that result must be communicated to parents to know the performance of their children through SMS service to their mobiles after every test session. Currently, there is no setup is available to analyze the results for the whole class / for one student to keep the record individually to focus on the performance to take steps for the improvement accordingly if required.
- Infrastructure (LED and Multimedia): The current infrastructure is not sufficient to teach the huge strength of class effectively stated by 39 female and 10 male respondents because Government of Province Punjab allowed only one (01) tablet for grade three (03) students in single school where teacher have to use it by taking from the head of the school and after due time, return back to the head of the school. It will be more effective for the whole class as well as teachers to use the big screen like LED or Multimedia to teach students the English language through the game / some interactive platform at the same time because currently, teacher has to call a student one by one manually to teach the English language on the tablet.

C. Students' Results

The findings collected under the SPSS (version 25.0) and provided in the form of a table and also represented after the students at public sector schools in Punjab were circulated by some teachers to their students to help address the questionnaire. Students ' demographic data indicates that the age of the respondents was only 06 (02%) in the 05-07 group, while the majority of the respondents were 217 (72.3%) in the 08-10 age group. 76 (25.3 percent), and finally only 01 (0.3 percent) in the age group of 11-13, belong to the age group of above 13. Of the 300 male participants, 138 were male which was 46% and the rest were female with 162, which corresponds to 54%, out of a total of 100%. Only 18 people who responded to the availability of ICT at home have their tablets at home, whereas the majority of respondents have 227 (75.9%) simple mobile phones at home, but 186 (62,2%) of respondents have smartphones at home for games and educational use. Only 16 (5.4 percent) and 100 (33.4 percent) interviewees have laptops and computers at home for games, cartoons or educational activities. In order to make ICT equipment open in class, 100% of respondents have tablets and computers for schooling, because the Punjab government has 01 tablets and 16 computers available in all public schools throughout the province, whereas the software is available for 9th and 10th-grade students only to teach guided reading.

In Table 13 139 (46.3 percent) respondents have, while 161 (53.7 percent) respondents do not have access to smartphones at home to play games and watch cartoons, knowing the frequency and proportion of the smartphones that they used to play games only, 138 (46.0 percent) of those surveyed, while 162 (54.0 percent) participants did not respond.

According to Table 14, Item 1 explains how LND software can be mastered only by 62 (20.7 percent) interviewees but the remainder of those respondents do not feel LND software is appropriate for English education because the overwhelming majority 283 (94.3 percent) of interviewees find LND software to be not useful since there is no integrated display and colors scheme while the main problem is the language used in the LND app which is not suitable according to 283 (94.3 percent) respondents, 280 of them (93.3 percent) answered that they cannot easily learn English using an LND application. Items 8 and 9 describe that vast majority 289 (96.3 percent) said they could not easily learn the comprehension by LND application, but 100.0 percent agreed that LND applications did not automatically give any content to learn comprehension and that 99.0 percent of the majority did not want to learn comprehension because of this.

In Table 15 (Point 11) it is defined that 100 (33.3%) of those approved, but 200 (66.7%) did not agree with the evaluation tool LND uses since 283 (94.3%) of those reacting to the assessment did not take part in enhancing comprehension of the English language. No application provides information to learn comprehension based on the results answered by 100.0% but 99.7% of respondents accepted to show the lack in understanding to the LND application.

Table 16 shows that 116 respondents say the LND application has no on-screen interactivity or buttons that attract respondents to use this application passionately to learn English, while the majority of respondents say 218 (72.7%) that contents and exercises are not available properly for the practice of English or animated cartoons. Similarly, students prefer the inclusion of pronunciation for better learning. Lastly, according to the students, content related to practicing and testing needs to be improved and enhanced.

TABLE. XIII. SMARTPHONE AND USAGE

Sr. #	**Item**	**Frequency**		**Percent**	
		Yes	**No**	**Yes**	**No**
1	Do you have an access to a Smartphone?	139	161	46.3	53.7
2	Do you use smartphone for playing games only?	138	162	46.0	54.0

TABLE. XIV. USABILITY, EASE OF USE OF LND APPLICATION

Sr. #	**Items**	**Frequency**		**Percent**	
		Yes	**No**	**Yes**	**No**
3	Do you think Literacy, Numeracy Drive (LND) software is suitable for you to learn English?	62	238	20.7	79.3
4	Do you think that LND application is interesting in learning English?	17	283	5.7	94.3
5	Do you think you can learn English easily from LND application?	20	280	6.7	93.3
6	Do you think that content of English learning is suitable in LND application?	17	283	5.7	94.3
7	Do you think teacher teaches English well with the help of LND application?	156	144	52.0	48.0
8	Do you think you can learn comprehension easily with the help of LND application?	11	289	3.7	96.3
9	Do you think application provides you content to learn comprehension automatically?	0	300	0.0	100.0
10	Do you think application develops interest to learn comprehension?	3	297	1.0	99.0

TABLE. XV. USEFULNESS OF LND

11	Do you think the method of assessment adopted in LND is suitable for you?	100	200	33.3	66.7
12	Assessment of LND improve your English learning?	17	283	5.7	94.3
13	Application provides the content to learn comprehension on the basis of results?	0	300	0.0	100.0
14	Application provides you your weaknesses to learn comprehension?	1	299	0.3	99.7

TABLE. XVI. FEATURES FOR IMPROVEMENT IN LND

		Responses	**Percent of Cases**
		N	
15	Do you think Interactive Screen and Icons are not available in LND application for you to learn English?	116	38.7%
16	Do you think Content and Exercises are not available in LND application for you to learn English?	218	72.7%
17	Do you think Animated Learning Cartoons are not available in LND application for you to learn English?	298	99.3%
18	Do you think Voice Pronunciation is not available in LND application for you to learn English?	291	97.0%

V. Conclusion

On the basis of result findings of questionnaire and interview, it is confirmed that LND mobile based application is not much efficient to teach students of Grade 03 because several problems identified specifically in English subject 1) Usability: this application is not usable because of its performance, non-availability of assistance, and not easy to use 2) Design: design is non-professional, not interactive color scheme of screens and icons, no proper layout, problem with font size, display of advertisements, no voice instructions, no animations, non-availability of updated questions 3) Content: the content in application is not aligned completely with curriculum, not suitable for teaching English, and application provider is not taking any steps to improve it 4) Accessibility: this application has not equal access and not equal time for every student to practice in due time of class because the of the huge strength of students 5) Infrastructure: the Government of Punjab allowed only 01 tablets for every school to use LND on it for teaching to grade 3 students which is not possible in huge strength of class with limited time of lecture, Finally 6) Assessment: the method of assessment is not suitable and the results are not communicated to parents as well as the results in LND does not show the status of class as a whole. No improvements are made on the basis or results because teachers have to cover up the syllabus of the book along with the practice of LND in 01 session of class.

VI. Recommendations

After conducting surveys and interviews from teachers and students to see the usefulness of LND in schools to teach students of grade 03, several problems were identified and the teacher also recommended some improvement for mobile-based LND applications to overcome all the issues.

- Game-Based learning should be adopted for improvement of user experience and the game's challenge must not have the issue of usability as it is necessary to keep the focus on usability while designing a game-based learning application. These points have also been highlighted by [21][23][58].

- The interface design should be attractive in game-based learning application that will have good impact [64] and encouraging the reading by playing games mentioned by [35] because many research studies illustrate that presentation of learning material through video games and animations inspire the students to gain cognitive skills and to improve comprehension. This is also mentioned by [15] in their study.

- The designers must be keeping in mind factors: usefulness, ease of use and clarity of goal that inspired the learners for using games to get the results better for effectiveness and encouraging behavior from learners by participating in such learning activities. These factors also highlighted in [59][73].

- Game-based learning should have music and pronunciation of words which helps the player to develop the motivation and also with the addition of feedback feature to improve the learning capability through a tactile observation of substances. This point is highlighted by [4].

- The game-based learning application should have targeted content because it could help more for the different context of learners and academics can achieve success than those who are failing to do so which is also mentioned by [36].

- The game-based learning must have fair assessment to measure the learning growth and also to keep track the gameplay because progress provides immediate feedback for teachers for their teaching objective and content which also described by [45] and [77] as well as surprises in the game that could be reason of inspiration for effective use of game also mentioned by [75].

Acknowledgment

This study is funded by the Universiti Kebangsaan Malaysia (The National University of Malaysia), Bangi, Selangor, Malaysia.

References

[1] Abdullah, F., & Ward, R. (2016). Developing a General Extended Technology Acceptance Model for E-Learning (GETAMEL) by analysing commonly used external factors. Computers in Human Behavior, 56, 238-256.

[2] Abubakar Muhammad, H. (2018). Teaching and Learning English Language in Nigerian Schools: Importance and Challenges. Teacher Education and Curriculum Studies, 3(1), 10. https://Doi.Org/10.11648/J.Tecs.20180301.13

[3] Ali, M. M., Malik, N. A., & Rehman, A. (2016). Mobile Assisted Language Learning (Mall) An Emerging Technology In English Language Class Rooms Of Lahore (Pakistan). Science International, 28(2), 9.

[4] Arnab, S., Petridis, P., Dunwell, I., & de Freitas, S. (2011). Enhancing learning in distributed virtual worlds through touch: a browser-based architecture for haptic interaction. In Serious Games and Edutainment Applications (pp. 149-167). Springer, London.

[5] Ashraf, H., Motlagh, F. G., & Salami, M. (2014). The Impact of Online Games on Learning English Vocabulary by Iranian (Low-intermediate) EFL Learners. Procedia - Social and Behavioral Sciences, 98, 286–291. https://doi.org/10.1016/j.sbspro.2014.03.418

[6] Bass, B. M., & Riggio, R. E. (2006). Transformational leadership. Psychology press.

[7] Behraam, S. (2015). Difficulties of Teaching English at Primary Level in Rural Areas of Pakistan, 02(04), 3.

[8] Bonilla Medina, S. X., & Cruz Arcila, F. (2013). Sociocultural factors involved in the teaching of English as foreign language in rural areas of Colombia: An analysis of the impact on teachers' professional development. Research in Teacher Education, 3(2), 28-33.

[9] Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M. & Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. Computers & Education, 94, 178-192.

[10] Brown,T. H., Mbati, L. S. (2015). Mobile learning: Moving past the myths and embracing the opportunities. International Review of Research in Open and Distributed Learning, 16(2), 115-135.

[11] Budiu, R. (2016, January 19). Mobile: Native Apps, Web Apps, and Hybrid Apps. Retrieved March 28, 2019, from https://www.nngroup.com/articles/mobile-native-apps/

[12] Cakir, I. (2006). The use of video as an audio-visual material in foreign language teaching classroom. Turkish Online Journal of Educational Technology-TOJET, 5(4), 67-72.

[13] Ching,K. Gwo,J. Chih,C. (2013). A personalized recommendation-based mobile learning approach to improving the reading performance of EFL students. Elsevier, Computers & Education, 53: 327–336.

[14] Coleman, H., & Capstick, A. (2012). Language in education in Pakistan: Recommendations for policy and practice. Islamabad: British Council.

[15] Conolly, T., Stansfield, M., Boyle, L. (2009): Games-Based Learning Advancements for Multy-Sensory Human Computer Interfaces: Techniques and Effective Practices. IGI Global Publishing.

[16] Crystal, D. (2012). English as a global language. Cambridge university press.

[17] Duolingo AI Research. (n.d.). Retrieved December 13, 2019, from https://ai.duolingo.com/.

[18] Elaish, M. M., Shuib, L., Abdul Ghani, N., Yadegaridehkordi, E., & Alaa, M. (2017). Mobile Learning for English Language Acquisition: Taxonomy, Challenges, and Recommendations. IEEE Access, 5, 19033–19047. https://doi.org/10.1109/ACCESS.2017.2749541.

[19] Emery, H. (2012). A global study of primary English teachers' qualifications, training and career development. 37.

[20] Fareed, D. M., Jawed, S., & Awan, S. (2018). Teaching English Language at SSC Level in Private Non-Elite Schools in Pakistan: Practices and Problems. Journal of Education and Educational Development, 5(1), 80. https://doi.org/10.22555/joeed.v5i1.1756.

[21] Gabriel, P., Hirashima, T., & Yusuke, H. (2018). A Serious Game for Improving Inferencing in the Presence of Foreign Language Unknown Words. International Journal of Advanced Computer Science and Applications, 9(2). https://doi.org/10.14569/IJACSA.2018.090202

[22] Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: an integrated model. MIS quarterly, 27(1), 51-90.

[23] Giessen, H. W. (2015). Serious Games Effects: An Overview. Procedia - Social and Behavioral Sciences, 174, 2240–2244. https://doi.org/10.1016/j.sbspro.2015.01.881

[24] Hamari, J., & Keronen, L. (2017). Why do people play games? A meta-analysis. International Journal of Information Management, 37(3), 125-141.

[25] Harper, C., & De Jong, E. (2004). Misconceptions about teaching English- language learners. Journal of adolescent & adult literacy, 48(2), 152-162.

[26] Hazaea, A. N., & Alzubi, A. A. (2018). Impact of Mobile Assisted Language Learning on Learner Autonomy in EFL Reading Context. Journal of Language and Education, 4(2), 48–58. https://doi.org/10.17323/2411-7390-2018-4-2-48-58

[27] Heil, C. R., Wu, J. S., Lee, J. J., & Schmidt, T. (2016). A review of mobile language learning applications: Trends, challenges, and opportunities. The EuroCALL Review, 24(2), 32-50.

[28] Hidayat, A., & Utomo, V. G. (2014). Open Source Based M-Learning Application for Supporting Distance Learning. Telkomnika, 12(3), 657.

[29] Huang, L-L., & Lin, C-C. (2011). EFL learners' reading on mobile phones. The JALT CALL Journal, 7(1), 61–78.

[30] Hussain, M. A., Niwaz, A., Zaman, A., Dahar, M. A., & Akhtar, M. S. (2010). Technology Based Learning Environment and Student Achievement in English as a Foreign Language in Pakistan. International Journal of Academic Research, 2(5).

[31] Ismail, H., Harous, S., & Belkhouche, B. (2016). A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis. Research in Computing Science, 110, 71-83.

[32] Karamouzian, F. M., Narcy-Combes, M.-F., & Ahmed, F. (2014). A Post-Use Evaluation of Pakistani Secondary School English Textbooks. 9(1), 24.

[33] Khan, T. J., & Khan, N. (2016). Obstacles in learning English as a second language among intermediate students of districts Mianwali and Bhakkar, Pakistan. Open Journal of Social Sciences, 4(02), 154.

[34] King, W. R., & He, J. (2006). A meta-analysis of the technology acceptance model. Information & management, 43(6), 740-755.

[35] Laamarti, F., Eid, M., & El Saddik, A. (2014). An Overview of Serious Games. International Journal of Computer Games Technology, 2014, 1–15. https://doi.org/10.1155/2014/358152

[36] Lameras, P., Arnab, S., Dunwell, I., Stewart, C., Clarke, S., & Petridis, P. (2017). Essential features of serious games design in higher education: Linking learning attributes to game mechanics. British journal of educational technology, 48(4), 972-994.

[37] Ley, B., Ogonowski, C., Hess, J., Reichling, T., Wan, L., & Wulf, V. (2014). Impacts of new technologies on media usage and social behaviour in domestic environments. Behaviour & Information Technology, 33(8), 815-828.

[38] Lim, C. K., Eng, L. S., Mohamed, A. R., & Mohamed Ismail, S. A. M. (2018). Relooking at the ESL Reading Comprehension Assessment for Malaysian Primary Schools. English Language Teaching, 11(7), 146. https://doi.org/10.5539/elt.v11n7p146

[39] Literacy & Numeracy. (2019). https://open.punjab.gov.pk/lnd/reports/open_slo_charts, Retrieved on December 24, 2019.

[40] Liu, M., Navarrete, C., Maradiegue, E., & Wivagg, J. (2014). Mobile learning and English language learners: A case study of using iPod touch as a teaching and learning tool. Journal of Interactive Learning Research, 25(3), 373-403.

[41] Marangunić, N., & Granić, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. Universal Access in the Information Society, 14(1), 81-95.

[42] Mchucha, I. R., Ismail, Z. I., & Tibok, R. P. Developing a Gamification-Based Thesaurus App to Improve English Language Vocabulary: A Case study of Undergraduate Students in Malaysia. CARNIVAL ON e-LEARNING (IUCEL) 2017, 43.

[43] McQuiggan, S., McQuiggan, J., Sabourin, J., & Kosturko, L. (2015). Mobile learning: A handbook for developers, educators, and learners. John Wiley & Sons.

[44] Miangah, T. M. (2012). Mobile-Assisted Language Learning. International Journal of Distributed and Parallel Systems, 3(1), 309–319. https://doi.org/10.5121/ijdps.2012.3126

[45] Minović, M., Milovanović, M., Šošević, U., & González, M. Á. C. (2015). Visualisation of student learning model in serious games. Computers in Human Behavior, 47, 98-107.

[46] Mohammad, N., Masum, R., Ali, Z., & Baksh, K. (2017). Teaching Practices of English Language in the Schools of Lasbela District, Pakistan. Case Studies, 6.

[47] Moon, J. W., & Kim, Y. G. (2001). Extending the TAM for a World-Wide-Web context. Information & management, 38(4), 217-230.

[48] Naheed, F. (2015). Interactive Approach in English Language Learning and Its Impact on Developing Language Skills. The Dialogue, (1), 9.

[49] National Education Policy Framework. (2018, November). Ministry of Federal Education and Professional Training. Pakistan.

[50] Park, S. Y. (2009). An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning. Educational technology & society, 12(3), 150-162.

[51] Petter, S., DeLone, W., & McLean, E. (2008). Measuring information systems success: models, dimensions, measures, and interrelationships. European journal of information systems, 17(3), 236-263.

[52] Pituch, K. A., & Lee, Y. K. (2006). The influence of system characteristics on e-learning use. Computers & Education, 47(2), 222-244.

[53] PMIU. (n.d.). Retrieved December 3, 2019, from https://open.punjab.gov.pk/lnd/reports/district_map

[54] Prensky, M. (2005). What can you learn from a cell phone? Almost anything! Innovate: Journal of Online Education, 1(5).

[55] Raja, R., & Nagasubramani, P. C. (2018). Impact of modern technology in education. Journal of Applied and Advanced Research, 3(S1), 33. https://doi.org/10.21839/jaar.2018.v3iS1.165

[56] Rasheed, S., Zeeshan, M., & Zaidi, N. A. (2017). Challenges of Teaching English Language in a Multilingual Setting: An Investigation at Government Girls Secondary Schools of Quetta, Baluchistan, Pakistan. International Journal of English Linguistics, 7(4), 149. https://doi.org/10.5539/ijel.v7n4p149

[57] Rashid, S., Cunningham, U., Watson, K., & Howard, J. (2018). Revisiting the digital divide(s): Technology-enhanced English language practices at a university in Pakistan. Australian Journal of Applied Linguistics, 1(2), 64–87. https://doi.org/10.29140/ajal.v1n2.7

[58] S. Deterding, (2015):The lens of intrinsic skill atoms: A method for gameful design, Human–Computer Interaction, vol. 30, no. 3-4, pp. 294–335.

[59] Schell, J. (2019). The Art of Game Design: A book of lenses. AK Peters/CRC Press.

[60] Shamim, F. (2008). Trends, issues and challenges in English language education in Pakistan. Asia Pacific Journal of Education, 28(3), 235-249.

[61] Somekh, B. (2008). Factors affecting teachers' pedagogical adoption of ICT. In International handbook of information technology in primary and secondary education (pp. 449-460). Springer, Boston, MA.

[62] Sönmez, A., Göçmez, L., Uygun, D., & Ataizi, M. (2018). A review of Current Studies of Mobile Learning. Journal of Educational Technology and Online Learning, 1(1), 12–27. https://doi.org/10.31681/jetol.378241

[63] Sun, H., & Zhang, P. (2006). The role of moderating factors in user technology acceptance. International journal of human-computer studies, 64(2), 53-78.

[64] The Foreign Language Center, Feng-Chia University, Taiwan, & Wang, B.-T. (2017). Designing Mobile Apps for English Vocabulary Learning. International Journal of Information and Education Technology, 7(4), 279–283. https://doi.org/10.18178/ijiet.2017.7.4.881

[65] Turan, B., Haşit, G. (2014). Teknoloji kabul modeli ve sınıf öğretmenleri üzerinde bir uygulama. Journal of Alanya Faculty of Business, 6 (1), 109-119.

[66] Turner, M., Kitchenham, B., Brereton, P., Charters, S., & Budgen, D. (2010). Does the technology acceptance model predict actual use? A systematic literature review. Information and software technology, 52(5), 463-479.

[67] UNESCO. (2003). Technology Education Guide. 153.

[68] Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. Information systems research, 11(4), 342-365.

[69] Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. Decision sciences, 39(2), 273-315.

[70] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS quarterly, 425-478.

[71] Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS quarterly, 36(1), 157-178.

[72] Viberg, O. (2015). Design and use of mobile technology in distance language education: Matching learning practices with technologies-in-practice. Repro: Örebro University.

[73] Wang, Y., Rajan, P., Sankar, C. S., & Raju, P. K. (2017). Let them play: the impact of mechanics and dynamics of a serious game on student perceptions of learning engagement. IEEE Transactions on Learning Technologies, 10(4), 514-525.

[74] Wiebe, E. N., Lamb, A., Hardy, M., & Sharek, D. (2014). Measuring engagement in video game-based environments: Investigation of the User Engagement Scale. Computers in Human Behavior, 32, 123-132.

[75] Wouters, P., van Oostendorp, H., ter Vrugte, J., Vandercruysse, S., de Jong, T., & Elen, J. (2017). The effect of surprising events in a serious game on learning mathematics. British journal of educational technology, 48(3), 860-877.

[76] Zhang, Y. (2016). The Impact of Mobile Learning on ESL Listening Comprehension. DEStech Transactions on Social Science, Education and Human Science, (icaem). https://doi.org/10.12783/dtssehs /icaem2016/4290

[77] Zhonggen, Y. (2019). A Meta-Analysis of Use of Serious Games in Education over a Decade. International Journal of Computer Games Technology, 2019, 1–8. https://doi.org/10.1155/2019/4797032

[78] Ishaq, K., Zin, N. A. M., Rosdi, F., Abid, A., & Farooq, U. (2019), Effectiveness of Literacy & Numeracy Drive (LND): A Students' Perspective.

[79] Hashim, H., Yunus, M. M., Embi, M. A., & Ozir, N. A. M. (2017). Mobile-assisted language learning (MALL) for ESL learners: A review of affordances and constraints. Sains Humanika, 9(1-5).

[80] Ishaq, K., Zin, N. A. M., Abid, A., Rosdi, F., & Ali, Q. (2020). Usefulness of Mobile Assisted Language Learning App. International Journal of Engineering and Advanced Technology (IJEAT), 9(3).

[81] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly, 319-340.

[82] Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management science, 46(2), 186-204.

Detecting Flooding Attacks in Communication Protocol of Industrial Control Systems

Rajesh L[1]
Research Scholar, Department of Electronics and Communications, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Vijayawada
Andhra Pradesh, India

Penke Satyanarayana[2]
Professor, Department of Electronics and Communications
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Vijayawada
Andhra Pradesh, India

***Abstract*—Industrial Control Systems (ICS) are normally using for monitoring and controlling various process plants like Oil & Gas refineries, Nuclear reactors, Power generation and transmission, various chemical plants etc., in the world. MODBUS is the most widely used communication protocol in these ICS systems, which is using for bi-directional data transfer of sensor data between data acquisition servers and Intelligent Electronic Devices (IED) like Programmable Logic Controllers (PLC) or Remote Telemetry Unit (RTU). The security of ICS systems is a major concern in safe and secure operations of these plants. This Modbus protocol is more vulnerable to cyber security attacks because security measures were not considered in mind at the time of protocol design. Denial-of-Service (DoS) attack or flooding attack is one of the prominent attacks for MODBUS, which affects the availability of the control system. In this paper, a new method was proposed, to detect user application-level flooding or DoS attacks and triggers alarm annunciator and displays suitable alarms in Supervisory Control and Data Acquisition system (SCADA) to draw the attention of administrators or engineers to take corrective action. This method detected highest percentage of attacks with less time compared to other methods. This method also considered all types of conditions, which triggers flooding attack in MODBUS protocol.**

Keywords*—Supervisory Control and Data Acquisition (SCADA); Remote Telemetry Unit (RTU); Programmable Logic Controllers (PLC); Communication Protocol; MODBUS; Industrial Control Systems (ICS)

I. Introduction

Industrial Control Systems (ICS) or Process Control Systems (PCS) are generally using for monitoring the field or process from a centralized location and control the field equipment to run the operations. Some of the examples of the industrial control systems using are oil and gas refineries, process and chemical plants, nuclear power plants, power generation and transportation [1]. The National Critical Infrastructure sectors are mainly depending on these ICS systems. The sectors which are very crucial for any country development in economy, social, technology are defined as National Critical Infrastructure [2].

Earlier these systems were confined to control room where all types of operations are taking place. But now-a-days these systems are connecting to internet, corporate networks to transfer the data to higher layer functionalities to meet corporate requirements like ERP, DMS and 3rd party services. Even though the connectivity of ICS systems to upper layers achieves the data sharing, but these systems are also opens doors to security attacks to destruct the functionality and country growth. It is required to protect these systems from cyber-attacks [3-4].

The Industrial Control Systems are connected to corporate networks and internet for sharing of SCADA data to 3rd party systems, taking remote for debugging and maintenance of the systems. This leads to security attacks and these systems are vulnerable to these cyber-attacks [5]. The Computer Emergency Response Team (CERT), an expert group that handles computer security incidents reports that the number cyber-attacks on ICS systems are increasing every year. These systems should be protected from security attacks for safe and secure operations of national critical infrastructures, where ICS systems play a vital role [6]. The number of cyber-attacks incidents during last 5 years was displayed in bar chart as shown in Fig. 1.

The security measures, which are suitable for Information and communications technology (ICT) systems, are not appropriate for ICS security because of their distinct purpose and functionality. There are areas of ICS systems where attacks may take place like network, computer hardware, controllers, interfaces etc [7]. The communication protocol, which is using for bi-directional data transfer between Data Acquisition Servers (DAQ) and Controllers, is one of the important areas where attacks are taking place. Modbus is most widely used communication protocol in ICS systems [8].

The field data from PLC will be transferred to SCADA Servers through communication protocols like Modbus, DNP etc. Modbus is a most widely used, open, application layer communication protocol for bidirectional data transfer between PLC and SCADA Servers. It is very simple and light weight communication protocol. It is based on simple request and reply message transfer [9]. The frame format of the Modbus protocol is shown in Figure 2. The SCADA Server sends the request to PLC and it will respond to the request and sends the response to the SCADA Server. If the request is valid then it will send valid response or if the request is not valid then it will respond with exception response [10]. The Modbus request frame contains device ID, function address, starting address, number of registers [11].

Modbus is lacking of security measures and suffering form number of security vulnerabilities. The Modbus protocol was designed without considering security in mind. There are number of vulnerabilities in Modbus protocol [12]. There is no checking of integrity, confidentiality, availability of this protocol [13]. The Modbus frame is very simple and knows to everyone because it is an open protocol. The frame is transferring in plain text without any encryption and checking integrity of the frame. There is no checking of authentication or authorization of master or target device. Any attacker which knows the IP address of PLC can send any command or any false command or response and can destroy the filed or process. Man-in-the middle attacks, replay attacks, Denial-of-Service (DoS) or flooding attack are some of the crucial attacks for Modbus protocol [12-13]. The Modbus is suffering from the following attacks [22]:

- There is no checking of authorization of source or target.
- There is no checking of authentication of connection.
- There is no checking of integrity of the frame. Anybody can change the content of the frame.
- The frame is transferring in plain text. Anybody can read the frame content.
- The attacker can seize the PLC.
- Replay attacks.
- Man-in-the middle attacks like false command injection, false response injection etc.

Denial-of-Service (DoS) or flooding attack is very crucial and had high impact on availability of the control system. The control system shall be available more than 99.95% for proper operations of the plant [14]. The PLC will be seized the control and cannot respond to the actual SCADA Server. The attacker can send malicious traffic to PLC and made the PLC busy with the flooding. The attacker can achieve this by sending SYNC packets continuously [15] or Internet Control Message Protocol (ICMP) packets or sending wrong Modbus requests at high rate.

In this paper, a new method was proposed to detect DoS or flooding attacks at user application level in Modbus protocol. Rest of the paper was described as follows: Section II describes the literature survey in this field. Section III explains the components of SCADA systems. The test set up used for this research was explained in Section IV. The proposed method, simulation of attacks and testing, detection of DoS attacks were described in Sections V & VI. The results of the testing and future work was explained and discussed in Section VII. The paper was concluded in Section VIII.

II. Literature Survey

The literature is available for security of industrial control systems and vulnerabilities, security attacks of Modbus protocol. Number of scholars worked for enhancing the security of Modbus protocol. Rajesh L et al. [12] described list of possible attacks and existing literature on security of ICS systems and Modbus protocol. Peter Huitsing et al. [13] list out the common security attacks in Modbus protocol. Rajesh L et. al [14] described the importance of ICS system availability and proposed method to enhance it. DoS attack seized the control of PLC and affects the system availability. Rajesh Kalluri et al. [15] explained simulation and impact of DoS attacks on SCADA systems. They did not propose any solution. Alvaro A. et al. [16] explained the various possible security vulnerabilities in industrial control systems and also described the real incidents where cyber-attacks took place across the world and mentioned the importance of security of these ICS systems. Jason Stamp et al. [17] also described the possible vulnerabilities in industrial control systems.

Fovino et al. [18] proposed a method for enhancing the security of Modbus protocol with AES, RSA algorithms. But the frame was transferred in plain text and attackers can read the message. Aamir Shahzad, S.Musa et al. [19] proposed algorithm for security in multicasting communication of Modbus protocol. This method did not consider the delay in communication response. It mentioned only number of detected attacks in percentage. The performance parameters are not compared and discussed. The research scholars worked and provided solutions for Man-in-the middle attack, replay attacks etc. [20].

The DoS attack was very less addressed. Bhatia et al. [21] simulated flooding attack by DoS by sending false commands. They proposed detection of flooding attacks based on anomaly detection and signature-based detection by Snort tool. They did not mention how many attacks are simulated and % of detection. They also not considered the DoS attack, which stops the services of PLC.

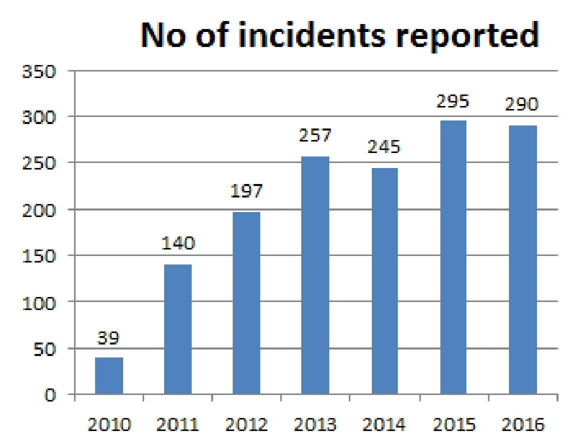


Fig. 1. Number of Incidents Reported by U.S. ICS-CERT (Ref: https://www.us-cert.gov/ics).

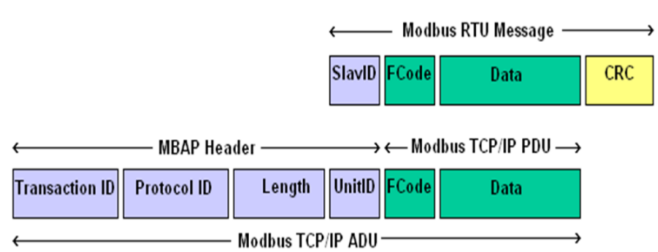


Fig. 2. MODBUS Frame Format.

From the literature review, it was concluded that Modbus protocol is suffering from security vulnerabilities, attackers could easily target this protocol. Some scholars provided solutions for some of these attacks, but DoS attack or flooding attack was not properly addressed for identification and detection, and it needs a solution or new method to properly detect the attack.

III. SCADA System Components

Any SCADA system mainly contains the following main components: [3].

- Sensors or field instruments.
- Programmable Logic Controllers (PLC) or Intelligent Electronics Devices (IED).
- Data Acquisition Servers, display work stations and other IT hardware components.
- Networking equipment.

A. Filed Instruments

Field instruments are basically transducers or sensors, used for measuring the field values like pressure, flow, density etc. These devices convert the physical quantity to electrical quantity and send the data to PLC/RTU. Monitoring and maintaining process variables at the appropriate levels is extremely critical in industrial automation and process control. A sensor in the industrial environment is either continuously or periodically measuring critical parameters such as density, temperature, pressure, flow, etc. The primary challenge of sensing in industrial environments is conditioning low signal levels in the presence of high noise and high-surge voltage.

B. RTU/PLC

Programmable Logic controller or Remote Telemetry Unit used for scanning the I/O and executing interlocks and logics for industry field operations. The basic units have a CPU (a computer processor) that is dedicated to run one program that monitors a series of different inputs and logically manipulates the outputs for the desired control. They are meant to be very flexible in how they can be programmed while also providing the advantages of high reliability compact and economical over traditional control systems. The I/O system provides the physical connection between the equipment and the RTU/PLC. The PLC/RTU will be connected to main SCADA Server though LAN or WAN. The PLC/RTU will have communication module for interfacing with SCADA Server through Serial or Ethernet communication.

C. DAQ Servers and IT Hardware

SCADA Server will be used for processing the received data from PLC/RTU and logging of the data for further future analysis. The Client will be used for display the data in different formats and sending the controls to PLC/RTU. SCADA package will be loaded in Server and protocol driver like Modbus, DNP will be running in this server. The main functionality of SCADA Servers is scanning the RTUs, time synchronization, database management, alarms triggering, report generation, control command execution etc.

D. Network Components

Network consists of Lan switches to connect the various nodes. Routers can also be used for WAN interface i.e to connect various stations. Redundancy is an important factor in SCADA networks. In pipeline applications the RTUs are geographically spread throughout the pipe line. Optical Fiber Communication will be used for bi-directional data transfer between main master station and RTUs.

The sensor data from field instruments will be collected by PLC or RTU. It will process the raw data and it transfers the data to centralized data acquisition system for further processing using suitable communication protocol like Modbus. Modbus protocol is suffering from security attacks and needs to protect the protocol. In next section, the test bench was described which is used in this research.

IV. Test Set up

A test bench was set up in our lab for performing the research and testing as shown in Fig. 3. Two computer systems, one number of Programmable Logic Controller (PLC) were connected in a local network through LAN switch. One of the computer systems were loaded with SCADA software and continuously polls the PLC to get the data or field values from the PLC. The Client component of Modbus protocol was loaded in SCADA Server and the server was polling the PLC through Modbus Protocol. The second system was loaded with Modbus simulator tool, which was used for simulating flooding attack for PLC in the network. The PLC was loaded with Modbus Server component, the developed software module to detect the attacks. The corresponding alarms were configured in SCADA MIMICs and trends. The suitable pop ups were created in SCADA system. The alarm annunciator was connected to PLC to trigger or energies the audible buzzer to draw the attention of engineers and operators in the control room.

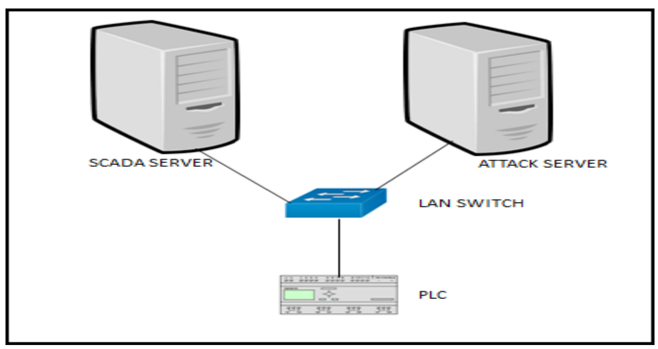


Fig. 3. Network Connectivity Diagram of Test Set up.

V. Proposed Method and Simulation of Flooding Attacks

The DoS attack by flooding PLC using Modbus protocol can be two types; first method is to stop the required services i.e., PLC not responding the legitimate requests from SCADA Server and other one is crashing the target and seizes the services i.e PLC was busy with responding attackers requests and denies the services from legitimate SCADA Server. It can be triggered by number of ways like sending more requests than pre-allocated maximum number of requests, sending wrong requests by making PLC busy, sending continuously SYNC packets or ICMP packets. In this set up the attack was triggered by sending the following cases:

1) By sending number of requests more than maximum pre-allocated in PLC Modbus driver.

2) By sending continuously correct Modbus requests at very high rate and.

3) By sending continuously wrong requests at high rate.

The above-mentioned attacks were simulated by second computer system, which was loaded with Modbus Master Simulator and configured to poll the PLC with arbitrary starting and number of registers continuously without any delay. The Modbus Master tool was configured with reading of coils, status, input registers and holding registers. Table-I displays the parameters configured in Modbus Slave tool.

TABLE. I. MODBUS REQUESTS SENDING TO PLC

Modbus ACTION	**Tag Parameters**		
	Function Code	***Starting address***	***Number of Registers***
Reading Coils	**1**	**1**	**10000**
Reading Status	**2**	**1**	**10000**
Reading Holding Registers	**3**	**1**	**2000**
Reading Input Registers	**4**	**1**	**2000**

An analog signal (PT) was simulated with ramp input in PLC logic and the same was received and plotted in SCADA mimic using Modbus protocol. The signal was started from 0 and incremented by 1 kg/cm2 for every one second in PLC memory registers. The simulation was done in PLC logic. The value was configured in trend for logging in SCADA system. The graph or trend plot was disturbed whenever the DoS attack was generated as shown in Fig. 4. Generally, the signal will be in saw tooth waveform if the system is continuously system available or non-presence of DoS attack. The PLC was not communicated with SCADA Server during DoS attack and the signal was not available at SCADA Server as well as mimic. In Fig. 4 it can be observed that the ramp signal is disturbed and it was flat for some time. The data is not available at SCADA Server during attack time interval because PLC was busy or not responding. DoS attack affects the system availability parameter.

The logic was also implemented for starting the pump to transfer the fluid between two Oil tanks. There are two tanks contains the Oil. The pump will start and the fluid will be transferring between tank A and tank B. Fig. 5 displays the flat curve of two tank levels after pump start and they did not resume the operation back. The PLC was not communicated back and it was out of control.

A digital output drive command was configured for start/stop the pump. Operator can give start/stop command from SCADA MMI to control the pump. The pump was started and the status was updated in SCADA. After launching of flooding attack, the stop command was sent. But PLC could not process the command because the PLC was busy in executing requests/commands from attacker. This was indicated in Fig. 6 In another experiment, the start command was sent and the pump was started at the field. The flooding attack was also triggered immediately after pump start command. The running status of pump was also reported to SCADA with delay of more than 5 sec. The same was observed in Fig.7.

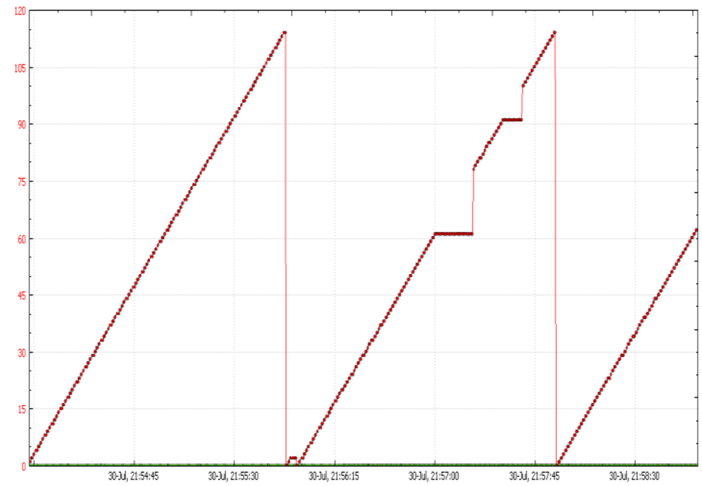


Fig. 4. DoS Attack Shown in SCADA Graph or Trend.

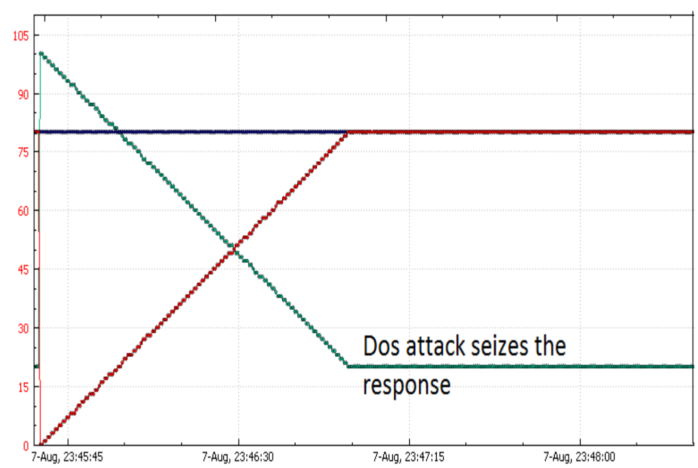


Fig. 5. DoS Attack Seized the Response from PLC.

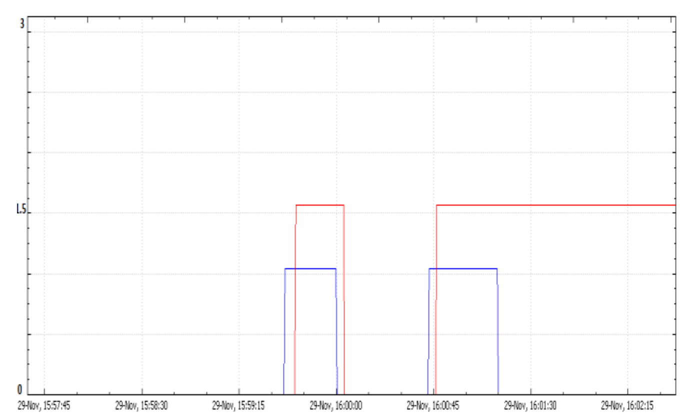


Fig. 6. Stop Command not Executed by PLC.

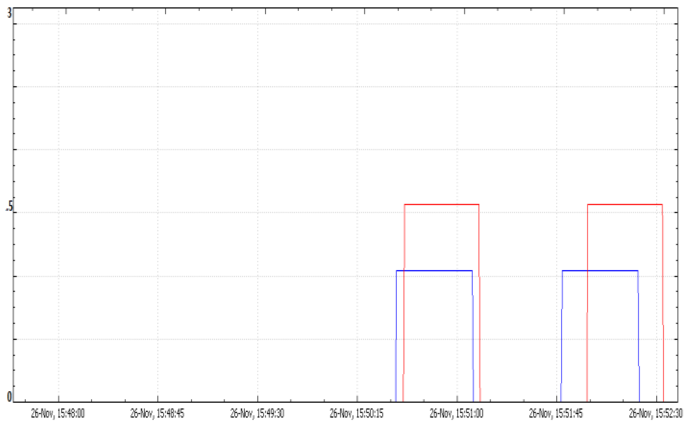


Fig. 7. Reporting of Pump Running Status at SCADA with Delay.

VI. Testing and Detection of DoS/Flooding Attacks

Whenever Modbus Master in SCADA system sent the Modbus request to PLC, the PLC saves the Modbus request parameters in registers or buffers based on these request parameters. SCADA server frames the Modbus requests and polls the PLC periodically. Therefore, PLC updates the starting address and number of registers to be polled, after one cycle of scanning. But it requires at least two cycles to update poll time interval. The buffer contains starting address, number of registers, function code, poll time interval and DoS attack recognition flag. Whenever the PLC receives the Modbus request from attackers, PLC checks the buffer values and detects that these values are not regular values and set the flooding attack recognition flag. If the flag is set, PLC drives a relay for annunciator, which is connected, to PLC. Whenever the PLC responds with exception response, the SCADA Server detects the same and received the flag and it triggers the pop up, generate alarm with audible sound to draw the attention of administrators or engineers that something was happened with PLC.

The PLC continuously monitor the Modbus request frame and stores the above request parameters like Function code, starting address, number of registers. Generally, the system was configured with required request blocks and the poll cycle is periodic with same parameters. The requests will be periodic with the same parameters for every scan. If any new request received with new parameters other than these stored parameters, PLC will set the DoS attack flag. If the PLC received wrong request and respond with exception response continuously more than three times, then also it sets the DoS attack flag. If the time duration between two consecutive Modbus requests is more than the saved values, then also it will trigger DoS flag. Generally, PLC will respond with correct response for Modbus request, but when it receives out of bound memory address or invalid memory address, it will send exception response. The suitable structure was defined in PLC for updating the parameters of Modbus requests.

The module also checks the IP address of incoming requests and filters which are not authorized. The PLC stores the authorized IP addresses and allows or passes the requests from the configured IP addresses. The module also filters the queries which are not configured in PLC memory by checking the incoming Modbus request parameters by saved or configured parameters.

The PLC also maintains the same parameters in its memory. Generally, once the system is commissioned and using for operations continuously, there will not be any changes in Modbus Configuration. For testing purpose, the above scenario was simulated by changing starting register and number of registers and the PLC detects that the starting and ending address was not matching with saved values from periodic cycle and sets the flag. Whenever the PLC detects any change in above parameters, it sets the DoS flag and the bit was received by SCADA also. Then SCADA will trigger a pop up with suitable message and audible sound, logs into day event reports and suitable alarm in alarm page.

VII. Results and Discussion

The developed module is useful for detecting DoS or flooding attacks at user application level. The attacker can make the PLC busy with sending wrong or correct Modbus requests to PLC. During the attack, PLC will not respond to legitimate server and data is not available in ICS system.

The DoS attack was detected as shown in Fig. 8 as per above developed module. Fig. 8 displays that DoS attacks received and alarm was triggered at SCADA level. The exception response or no response was received from PLC in SCADA during the attack. The PLC detects the DoS attack as per our developed module, sets a flag in its memory, and can be transferred to Modbus. The SCADA system detects the flag after re-establish the communication with PLC and triggers the pop ups and alarms at SCADA level.

```
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0
Recv EXCEPTION CODEs=84  TID=0

time for total db ET=860163664 ST=8

 ++no of request sent=407 no of co

 ALARM FOR DoS ATTACK TRIGGERED
```

Fig. 8. DoS Attack was Detected in SCADA.

The proposed method was tested for 100 times by simulating DoS attacks by sending the following requests indicated in table 2, continuously as per the table and successfully detect 98 attacks within 5 sec. Two instances of Modbus Clients were configured and connected to PLC. Client 1 was configured with the parameters as shown table and polls for every 10 msec. Client 2 was configured with different parameters and polls for every 1 msec. Table 2 shows one instance of Modbus request parameters for client 1 and client 2. The parameters were changed for every trail and results are recorded. The method was successfully detected 98% attacks. This method offers highest % of detection of attacks with less time compared to other methods. This method also simulated all types of conditions by which flooding attack or DoS attack can trigger. Audible buzzer was energized whenever the attack was detected. Suitable pop ups and alarms were also displayed in SCADA mimic.

TABLE. II. MODBUS REQUESTS SENT TO PLC BY MODBUS CLIENTS

MODBUS CLIENT-1	**For every 10 msec**		
	Function Code	***Starting address***	***Number of Registers***
MODBUS requests by Modbus Client 1	**1**	**1**	**10000**
	2	**1**	**10000**
	3	**1**	**2000**
	4	**1**	**2000**
MODBUS CLIENT-2	**For every 1 msec**		
	Function Code	***Starting address***	***Number of Registers***
MODBUS requests by Modbus Client 2	**1**	**5000**	**5000**
	2	**5000**	**5000**
	3	**2000**	**5000**
	4	**2000**	**5000**

VIII. CONCLUSION AND FUTURE WORK

Industrial Control Systems plays a vital role in National Critical Infrastructures. MODBUS Protocol is most widely used communication protocol in industrial control systems. The protocol is more vulnerable to security attacks and it needs to protect the system. Modbus is prone to number of security attacks because it was designed without security measures. Daniel of Service (DoS) attack is one of the important attacks which affect the availability of the control system. In this paper, a different method was proposed to detect application level flooding or DoS attack in Modbus protocol. In this test set up the DoS attack was simulated at application level by sending correct and wrong Modbus requests at very high rate and sending Modbus requests more than configured maximum number of Modbus counter value and the same was detected by PLC and SCADA systems. The attacks were simulated with various parameters of Modbus requests and 98% attacks were detected successfully. In future, we will work on developing solutions to address other cyber-security attacks of Modbus Protocol.

REFERENCES

[1] Patrick, Dale R., and Stephen W. Fardo. Industrial process control systems. The Fairmont Press, Inc., 2009.

[2] Alvaro A. C ardenas, Saurabh Amin, Shankar Sastry. "Research Challenges for the Security of Control Systems", 3rd USENIX workshop on Hot Topics in Security (HotSec '08). Associated with the 17th USENIX Security Symposium. San Jose, CA, USA. July 2008.

[3] Sullivan D., Luiijf E., Colbert E.J.M. (2016) "Cyber-security of SCADA and Other Industrial Control Systems, Components of Industrial Control Systems", Advances in Information Security, vol 66. Springer, Cham.

[4] Figueroa-Lorenzo, Santiago, Javier Añorga, and Saioa Arrizabalaga. "A Role-Based Access Control Model in Modbus SCADA Systems. A Centralized Model Approach." Sensors 19.20 (2019): 4455.

[5] Alguliyev, Rasim, Yadigar Imamverdiyev, and Lyudmila Sukhostat. "Cyber-physical systems and their security issues." Computers in Industry 100 (2018): 212-223.

[6] Nazir, Sajid, Shushma Patel, and Dilip Patel. "Assessing and augmenting SCADA cyber security: A survey of techniques." Computers & Security 70 (2017): 436-454.

[7] Coffey, Kyle, et al. "Vulnerability assessment of cyber security for SCADA systems." Guide to Vulnerability Analysis for Computer Networks and Systems. Springer, Cham, 2018. 59-80.

[8] Nardone, Roberto, Ricardo J. Rodríguez, and Stefano Marrone. "Formal security assessment of Modbus protocol." 2016 11th International Conference for Internet Technology and Secured Transactions (ICITST). IEEE, 2016.

[9] MODBUS Over Serial Line Specification & Implementation Guide V1.02, Modbus Organization, Dec 20, 2006.

[10] MODBUS Messaging On Tcp/Ip Implementation Guide V1.0b, Modbus Organization, Oct 24, 2006.

[11] MODBUS Appl Protocol Specification V1.1 b3, Modbus Organization, April 26, 2012.

[12] Rajesh, L & Satyanarayana, P., "Communication protocol security in industrial control systems to protect national critical infrastructure". Journal of Advanced Research in Dynamical and Control Systems. 9. 290-304, 2017.

[13] Huitsing, Peter, Rodrigo Chandia, Mauricio Papa, and Sujeet Shenoi. "Attack taxonomies for the Modbus protocols." International Journal of Critical Infrastructure Protection 1 (2008): 37-44. doi:10.1016/j.ijcip.2008.08.003.

[14] Rajesh, L & Satyanarayana, P, "Dual channel scanning in communication protocol in industrial control systems for high availability of the system", International Journal of Technical and Physical Problems of Engineering. Vol. 11 P.22-27, 2019.

[15] Kalluri, Rajesh & Lagineni, Mahendra & Kumar, R. & Prasad, G.L., "Simulation and impact analysis of denial-of-service attacks on power SCADA". 1-5. 10.1109/NPSC.2016.7858908.,2016.

[16] Alvaro A. C ardenas, Saurabh Amin, Shankar Sastry. "Research Challenges for the Security of Control Systems", 3rd USENIX workshop on Hot Topics in Security (HotSec '08). Associated with the 17th USENIX Security Symposium. San Jose, CA, USA. July 2008.

[17] Jason Stamp, John Dillinger, William Young, and Jennifer DePoy., "Common vulnerabilities in critical infrastructure control systems" SANS SANSFIRE 2003 and National Information Assurance Leadership Conference V – (NIAL), July 14-22, 2003, Washingdon, DC

[18] Fovino, I. N., Carcano, A., Masera, M., & Trombetta, A., "Design and Implementation of a Secure Modbus Protocol"., IFIP Advances in Information and Communication Technology Critical Infrastructure Protection III, 83-96. doi:10.1007/978-3-642-04798-5_6, 2009.

[19] Shahzad, A.A. and S. Musa, "Cryptography and authentication placement to provide secure channel for SCADA communication". International Journal of Security., 6: 28-44, 2015.

[20] Shahzad, A., Lee, M., Lee, Y., Kim, S., Xiong, N., Choi, J., & Cho, Y.,"Real Time MODBUS Transmissions and Cryptography Security Designs and Enhancements of Protocol Sensitive Information". Symmetry,7(3), 1176-1210. doi:10.3390/sym7031176, 2015.

[21] S. Bhatia, N. Kush, C. Djamaludin, J. Akande, E. Foo, "Practical Modbus Flooding Attack and Detection", Proceedings of the Twelfth Australasian Information Security Conference (AISC) ser. AISC'14, vol. 149, pp. 57-65, 2014.

[22] Luo, Xuan, and Yongzhong Li. "Security Enhancement Mechanism of Modbus TCP Protocol." DEStech Transactions on Computer Science and Engineering iciti (2018).

An Artificial Deep Neural Network for the Binary Classification of Network Traffic

Shubair A. Abdullah[1]
College of Education, Department of Instructional and learning Technology, Sultan Qaboos University
Muscat, Oman

Ahmed Al-Ashoor[2]
College of Pharmacy, Department of Pharmacognosy
University of Basra
Basra, Iraq

***Abstract*—Classifying network packets is crucial in intrusion detection. As intrusion detection systems are the primary defense of the infrastructure of networks, they need to adapt to the exponential increase in threats. Despite the fact that many machine learning techniques have been devised by researchers, this research area is still far from finding perfect systems with high malicious packet detection accuracy. Deep learning is a subset of machine learning and aims to mimic the workings of the human brain in processing data for use in decision-making. It has already shown excellent capabilities in dealing with many real-world problems such as facial recognition and intelligent transportation systems. This paper develops an artificial deep neural network to detect malicious packets in network traffic. The artificial deep neural network is built carefully and gradually to confirm the optimum number of input and output neurons and the learning mechanism inside hidden layers. The performance is analyzed by carrying out several experiments on real-world open source traffic datasets using well-known classification metrics. The experiments have shown promising results for real-world application in the binary classification of network traffic.**

Keywords—Deep learning; ANN; packet classification; binary classification; malicious traffic classification

I. Introduction

The classification of network packets refers to the task of identifying abnormal behavior in networks. Currently, governmental and organizational networks across the world are natural targets for attackers who aim to compromise them in order to perform illegal activities such as information stealing. As each generation of malware is progressively more advanced, the development of successful online intrusion detection systems is at the forefront of information security tasks.

Naturally, a host compromised by malware will most probably generate packets that serve the malware's activities, i.e. malicious packets. A packet is a container used to carry data over a network. It normally represents the smallest amount of data that can traverse over a network at a single time. Normal TCP/IP packets contain several forms of information, including the data it is carrying, source and destination IP addresses, source and destination port numbers, and other information related to the quality of service and packet handling. A straightforward way to detect intrusions is with packet classification, which could be implemented using machine learning techniques. Machine learning is an application of artificial intelligence that provides software with the ability to automatically learn and evolve from experience without being explicitly programmed. It could be used to solve problems of predictions and classifications. In general, machine learning techniques are divided into two types: supervised and unsupervised. Packet classification is modeled as classification problem in supervised learning. Supervised learning has a set of input features and output classes. It has an algorithm to learn the mapping function from the input features to the output class. The goal is to approximate the mapping function. When new input features are introduced, the algorithm predicts the output class. In unsupervised learning, on the other hand, there is a set of input features without corresponding output classes to perform the learning task. The goal for unsupervised learning is to model the underlying distribution in input data to learn more about the data [1].

Deep learning is a subset of machine learning. In a deep learning system, multiple layers, i.e. input, hidden, and output layers, are stacked to form a neural network. Each hidden layer applies neuron mathematical structures to perform the learning task. The learning approach is designed to analyze data continually with a logic structure similar to how a human would draw a conclusion. The data analysis is repeated as long as inaccurate predictions occur. When the system returns a low accurate prediction level, the learning approach will automatically make an adjustment. Usually, a deep learning neural network has more than one hidden layer, which determines the network depth between the input and output layers. The learning process consists of two crucial elements: forward feature abstraction and backward error feedback. The first element is important for input data analysis and the second is important for tweaking the neurons [2].

Two gaps were observed in the literature related to the classification of the network packet problem, though there is a long record of research in packet classification over the last three decades [3]. The first gap is that the research field is still far from finding a perfect system with high malicious packet detection accuracy. The second gap is the lack of comprehensive research attempts that have employed deep learning approaches to classify network packets. Since it is a relatively new research area, there have been few research attempts investigating, evaluating, and tuning well-known deep learning approaches to classify the network packet problem [4].

The main contribution of this paper is to fill the abovementioned gaps by designing and implementing an artificial deep neural network (ADNN) using the state-of-the-art methodologies of deep neural networks. The ADNN is evaluated using standard classification quality metrics and compared with well-known classification algorithms including kNN, SVM, and Naive Bayes.

The rest of the article is divided into five sections. Section II reviews notable research in the area. Section III provides the architectural designs of the ADNN proposed in this study along with the implementation details, including the hardware and software used. Section IV presents the experimental results. Finally, Section V concludes the paper and presents some future facts.

II. Related Work

Several research papers have been published in the last decades dealing with enhancing the performance of network packet classifications. Most published papers have employed both supervised and unsupervised machine learning approaches. Examples of supervised approaches employed include support vector machine (SVM) [5] and k-nearest neighbors (kNN) [6]. For unsupervised approaches, the most common employed approach is k-means clustering [7]. Interested readers may refer to the work of Nguyen et al. [8] and Dainotti et al. [9] for a detailed overview of the machine learning techniques applied to traffic classification. Abdullah et al. [10] proposed a novel evolving fuzzy system to discriminate anomalies by inspecting the network traffic. The system incorporated the knowledge base-evolving mechanism and showed a significant positive impact on the classification accuracy. An open source tool for network traffic classification called the traffic identification engine (TIE) was developed in 2008 and gradually evolved over the years from 2009 to 2014 through the support of the open source community. TIE uses a combination of different traffic classification techniques and can be applied to both live traffic and previously captured traffic traces [11].

The application of deep neural network approaches such as deep autoencoders, deep belief neural (DBN) networks, deep convolutional neural networks (CNN), and recurrent neural networks (RNN) to solve the packet classification problem is a relatively new area of research. These approaches have already shown excellent capabilities in dealing with real-world problems such as facial recognition [12], intelligent transportation systems [13], etc. Lotfollahi et al. [14] proposed the "deep packet" system employing the deep CNN approach to integrate feature extraction and classification. Deep packet can handle traffic characterization to categorize network traffic into classes, i.e. FTP and P2P, and application identification to identify end-user application e.g. BitTorrent and Skype. Rahul [15] applied deep learning techniques to the classification of network protocols and applications using flow features and data signatures. They used their own dataset for traffic identification and the Microsoft Kaggle dataset for malware classification tasks. The DBN network is a type of generative neural network that uses an unsupervised machine learning model to produce results. Alom [16] explored the capabilities of the DBN network in performing intrusion detection. They performed a series of experiments after training the DBN network with the NSL-KDD dataset. The RNNs are designed for sequence prediction problems, which involve using historical sequence information to predict the next values or next single value in a sequence. Lopez-Martin et al. [17] presented a complete study on several architectures that integrate a CNN and an RNN. They showed that the integration of RNN with CNN could provide the best results for the Internet of Things (IoT) network traffic classification.

Despite these efforts, the literature lacks comprehensive attempts that have investigated, evaluated, and tuned well-known deep learning approaches for classifying network packets. The strategy used in the research methodology in this paper is to investigate and experiment each stage in building the network separately. Moreover, the final stage involves tuning the parameters in order to reach the highest possible level of accuracy in classifying network packets into malicious and normal packets.

III. Methodology

A. Dataset Description

Suppose that D is a supervised training dataset for network packet classification with i-tuple elements. D is divided into two subsets: DN contains normal packets and DM contains malicious packets:

$$D_N \subset D \wedge D_M \subset D \leftrightarrow D \equiv D_N \cup D_M$$

The D set can be represented by the set builder notation:

$$D = \{x | x \in D_N \vee x \in D_M\}$$

Where x is 11-tuple element that includes 10 features plus 1 class for describing the packet. The class of the packet c is defined as follows:

$$c = \begin{cases} 0 \; if \; x \in D_N \\ 1 \; if \; x \in D_M \end{cases}$$

In a normal situation, i.e. where there are no malicious packets, packets are considered as normal, that is $X_N = \{x_1, x_2, x_3, x_n\} \subset D_N$. This situation can be represented as follows:

Let P(x) denote $x \in X_N$ where $X_N \subset D_N$

Then the truth-value of $\forall \, x \, P(x)$ is True (1)

The truth-value of (1) is changed to false if the universe of discourse contains normal packets and malicious packets, that is $X_{NM} = \{x_1, x_2, x_3, x_n\}$, $X_{NM} \subset D_N \wedge X_{NM} \subset D_M$. This new situation is represented as follows:

Let P(x) denote $x \in X_{NM}$ where $X_{NM} \subset D_N \cup D_M$

Then the truth-value of $\forall \, x \, P(x)$ is False (2)

The truth-value of $\exists x \, P(x)$ is True (3)

In this case x is called a counterexample for (2) since it turns its truth-value into false. The objective of the ADNN is to identify and classify the counterexamples as malicious.

The dataset used in this research was prepared from the UNSW-NB 15 dataset, which has been created by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) in the University of New South Wales, Canberra, Australia [18]. It contains a hybrid of real modern normal activities and synthetic contemporary attack behaviors. The training set contains 175,341 records and the testing set contains 82,332 records from both normal and malicious packets. For each record, there are 48 features (dependent variables) and one label (dependent variable). The preparation of training and test datasets involved.

1) Feature selection: this task aims to find and select the most useful features in a dataset. The features with low importance are removed. For example, the feature *swin*, which refers to the value of source TCP window advertisements and the feature *dwin*, which refers to the value of destination TCP window advertisements.

2) Encoding categorical features: this task aims to convert categorical features into numeric values. Three categorical features exist in the dataset: proto, service, and state. The values in these features have no ordinal relationship. Therefore, integer encoding was used [2]. Each unique category value was assigned an integer value. Table I describes the features that are selected.

3) Dataset filtration: this task is done by removing the records with a high percentage of missing values. For example, any record with a service marked as (- hyphen) or with duration equal to zero was removed from the dataset. Table II describes the dataset after completing the filtration task.

4) Feature scaling: this task is important when working with a learning model. It aims to scale the features to a range centered on zero. It prevents features that have high variance from dominating other features in the dataset. The standard scaler and the results were as expected and the features are normalized so that they have mean = zero and standard deviation = one.

B. *ADNN Architecture*

The architecture of the ADNN is shown in Fig 1. It is composed of four layers, namely an input layer, two hidden layers, and an output layer.

TABLE. I. DESCRIPTION OF 12-TUPLE ELEMENTS (PACKET FEATURES)

#	Feature	Description
1	dur	Record total duration
2	proto	Transaction protocol (TCP | UDP)
3	service	0=http, 1=ftp, 2=smtp, 3=ssh, 4=dns, 5=ftp-data , 6=irc and 7=(-) if not a much used service
4	state	Indicates the state and its dependent protocol, e.g. 0=ACC, 1=CLO, 2=CON, 3=ECO, 4=FIN, 5=REQ, and 6=RST
5	spkts	Source to destination packet count
6	sbytes	Source to destination transaction bytes
7	sttl	Source to destination time to live value
8	sload	Source bits per second
9	swin	Source TCP window advertisement value
10	synack	TCP connection setup time, the time between the SYN and the SYN_ACK packets
11	ackdat	TCP connection setup time, the time between the SYN_ACK and the ACK packets
12	label	0= normal and 1=malicious

TABLE. II. DATASET DESCRIPTION

Item	Description
Total number of records	210,191 records
Normal packets	124,709 records
Malicious packets	85,482 records

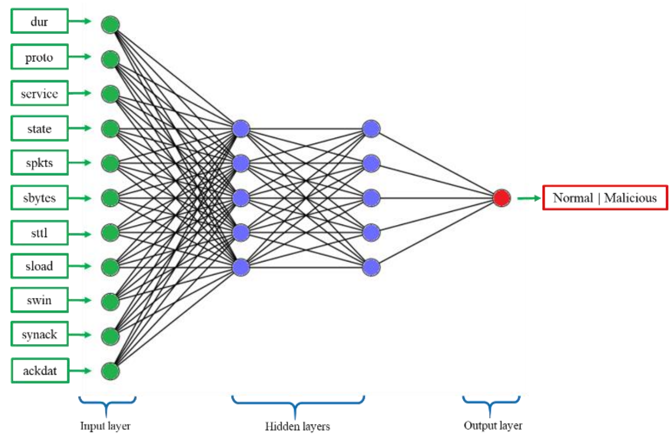


Fig. 1. ADNN Architecture.

The input layer is the first layer of the ADNN. It does not apply any operations and has no associated values of weights. It consists of 11 neurons, one neuron for each input feature. Given a set of training samples $\{x_1, x_2, x_3, \ldots x_n\}$, where $x_i \in X_N \vee x_i \in X_M$, an input neuron accepts x_i and passes it to one or more neurons in the next layer - the first hidden layer.

Two hidden layers were created in the ADNN and each layer contains five neurons. All the neurons are connected to every neuron in the next layer. For each neuron, there are a certain number of inputs and weights. The number of weights for a neuron equals the number of its input values. Each neuron in hidden layer #1 has 11 inputs and 11 weights, and each neuron in hidden layer #2 has five inputs and five weights. Weights are crucial to ADNN functioning because they are learnable parameters. The values of weights are initialized randomly to be close to zero but not zero before the learning starts. When presented with data during training, their values are adjusted to new values, and this adjustment will contribute to deciding the importance of inputs.

Three operations are done by a single neuron. First, it calculates the weighted summation of all the input values (x_n). Then, it applies an activation function to the weighted summation. Finally, it passes the results to a neuron in the next layer, as shown in:

$$\acute{Z} = \sum_{i=1}^{m} (w_i x_i) \tag{4}$$

$$\bar{O} = \varnothing(\acute{Z}) \tag{5}$$

Where w_i is an input data x_i weight, m is the number of neuron input data, $\acute{Z}$ is the weighted summation, $\bar{O}$ is the output of the neuron, and Ø *(theta)* is the activation function.

The activation function is responsible for transforming the weighted summation from the neuron into the activation of the next neuron. There are several activation functions in the literature. In this study, we used the rectified linear activation unit, or ReLU for short, for two reasons: (1) its computational simplicity and (2) its linear behavior increases the chances of optimizing the ADNN [19]. The ReLU activation function is formalized as below:

$$\emptyset(\acute{Z}) = max(\acute{Z}, 0) \quad (6)$$

This is the last layer in the ADNN. It receives input from hidden layer #2, makes some transformation, and outputs a binary (zero = normal or one = malicious). It consists of a single neuron that calculates the weighted summation of its input values and applies the sigmoid activation function to produce the final output. As we have two events that are mutually exclusive and cannot both occur at the same time (normal traffic and malicious traffic), we used the sigmoid activation function, which performs perfectly in this type of classification problem. Moreover, a single sigmoid neuron can be used to estimate the probability p(y=1) [19]. The sigmoid activation function is represented as below:

$$\bar{y} = \emptyset_1(\acute{Z}) = \frac{1}{1+e^{-\acute{Z}}} \quad (7)$$

Where $\acute{Z}$ is the output of hidden layer #2 calculated as in (5) and (6), $\bar{y}$ is the output of the neuron, and $\emptyset_1$ is the sigmoid activation function. Fig. 2 illustrates the architecture of the hidden and output layers.

The input $\{x_1, x_2, x_3, \ldots x_n\}$, where $x_i \in X_N \vee x_i \in X_M$ provides the initial information that propagates to the hidden neurons at each layer and finally produces the output $\bar{y}$, which is a number in the range from 0 – 1. We used the cross-entropy loss function [2] to compute the average error across all examples. The cross-entropy loss function is represented as follows:

$$H(y, \bar{y}) = -\sum_{i=1}^{n} y_i \log \bar{y}_i$$

Where y is the actual value, $\bar{y}$ is the output of the ADNN, and $H(y, \bar{y})$ is the cross-entropy loss function. After each forward propagation, the ADNN seeks a set of weights that minimize the difference between $\bar{y}$ and y. To get the least possible difference, the ADNN backpropagates the information about the error through the layers in order to tweak the weights and recalculate a new $\bar{y}$. We used the adaptive moment estimation (Adam) optimizer [20], which is a search technique to tweak weights in each neuron in the hidden layers. Adam is an adaptive learning rate optimizer that has been designed specifically for training deep neural networks. There are other options for optimizing the weights of neurons i.e. root mean square propagation (rmsprop), which is a gradient-based optimization technique.

In deep learning, when an entire dataset is passed forward and backward through the neural network once, this full cycle is called an epoch. The number of epochs is a tunable parameter, and usually more than one epoch is used. To optimize the learning, we used 20 epochs to train the ADNN. The batch size, which is the number of training examples in one epoch, is set to 10 samples in order to avoid overloading the processor and the RAM of the computer.

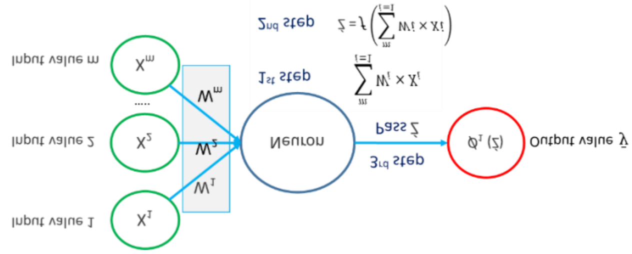


Fig. 2. Neuron of Hidden and Output Layers.

IV. Experiments

The experiments conducted aligned with the strategy of the research methodology. Five experiments were conducted aiming to gradually and systematically build and optimally set up the ADNN. At each experiments, an investigation task for a deep learning technique was performed.

Prominent metrics were used to evaluate the classification quality of the ADNN, such as accuracy, area under curve (AUC), recall, precision, and F1. These evaluation metrics were computed using a confusion matrix, which presents four measures: True Positive (TP): malicious traffic is classified by the ADNN as malicious traffic; False Positive (FP): normal traffic is classified by the ADNN as malicious traffic; True Negative (TN): malicious traffic is classified by the ADNN as normal traffic; False Negative (FN): normal traffic is classified by the ADNN as normal traffic.

A. Initial Experiment

The initial experiment was conducted in a straightforward way only to verify the code implementation and the parameter configurations. The dataset was split randomly into 75% training set (157,643 samples) and 25% test set (52,548 samples). The values for the number of epochs, batch size, and optimizer are 20, 10, and 'Adam' respectively. Fig. 3 shows the results of fitting the ADNN to the training set. It shows the accuracy for each epoch. The accuracy of the first epoch is 81%, which then increases steadily until it reaches a peak of 86% in the eleventh epoch. The accuracy begins to stabilize at slightly below 85% in the fourteenth epoch. The mean value of 84% along with the variance 0.01 show that the accuracies of 20 epochs are related to each other.

Figure 4 shows the results of classifying the test set in terms of the confusion matrix. The accuracy is calculated using equation, and te result generated is 84%:

$$accuracy = (tp + tn)/(tp + tn + fn + fp)$$

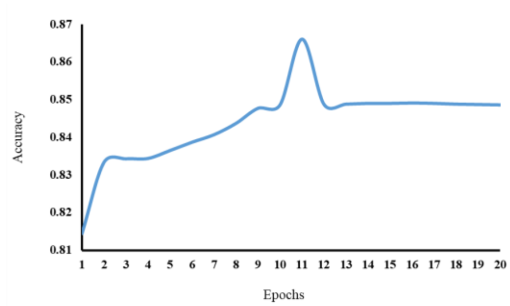


Fig. 3. Epoch Accuracies for Fitting the Training Set.

	Positive	Negative
Positive	TP=31,212	FN=941
Negative	FP=7,291	TN=13,104

Fig. 4. Confusion Matrix for Classifying the Test Set in the First Experiment.

B. K-Fold Cross-Validation Technique

In the previous experiment, the ADNN was trained using a 75% training set (157,643 samples), with the result revealing a low accuracy both for the training set and test set. Judging the ADNN performance on the accuracy obtained from one test set does not give a complete idea of the performance with regard to variance. Variance occurs when very different accuracies are obtained after testing a model using different test sets. In order to optimize the method used to evaluate the ADNN, the k-fold cross-validation technique is employed in this experiment. The advantage of this technique is that all samples are used for both training and validation, with every single sample being used for validation exactly once. The following steps were followed in this experiment:

1) The original dataset comprised of 210,191 samples was randomly partitioned into 10 equal sized subsets. Each subset contained 12,019 samples. k=10 was chosen as it is commonly used in the literature.

2) The partitioning of the original dataset into 10 subsets was governed by criteria to ensure that each subset has 60% normal samples and 40% malicious samples. A stratified cross-validation process that is common variation of cross-validation to ensure each subset has the same proportion of normal and malicious samples was used. We used a 60/40 proportion to create a semi-stratified cross-validation.

3) Of the 10 subsets, a single subset was used to testing. The remaining 9 subsets were used as the training sets.

4) The cross-validation process was repeated 10 times, with each subset being used only one time as the test set.

5) The values of the epochs and batch size variables used in the previous experiment were used again.

To calculate the accuracy, the 10 results were averaged. Fig. 5 shows the results of the semi-stratified 10-fold cross-validation experiment. The total number of samples was 210,191. In each of the 10 folds, there were 189,172 samples as the training set and 21,019 samples as the test set. In contrast with the accuracy of 84% obtained in the training phase of the previous experiment, the mean accuracy of 86% reflects an improvement in the building of the ADNN. The resulting low variance of 0.004 also suggests an improvement in the ADNN.

To determine the accuracy of the ADNN precisely, a test set of 50,000 unseen samples was prepared to test the ADNN performance on unseen samples. Fig. 6 shows the results of classifying the test set in terms of the confusion matrix. An accuracy of 84% was calculated - the same value calculated in the previous experiment. Although the variance obtained in the training phase is quite low, the accuracy obtained from the test phase indicates the presence of a bias. The low accuracy means that there is a difference between the average prediction of the ADNN and the correct value.

C. Dropout Technique

The challenge was to beat the low accuracy of 84% obtained from testing the ADNN on the unseen 50,000 samples in the previous experiment. The accuracy obtained from the training part of 86% was probably the result of using a small dataset, which may cause overfitting and poor performance. When the ADNN was faced with the unseen 50,000 samples, it predicted them with lower accuracy than in the training. In such a situation, there is a need for regularization. Dropout is an approach to regularize deep neural networks that helps reducing interdependent learning amongst the neurons [21]. It refers to dropping out randomly selected neurons from a certain layer during the training. Consequently, the outputs of the dropped neurons are not considered during a particular forward or backward pass. Normally, the dropout technique is applied on the hidden layers and has been proven to enhance the performance of deep neural networks over other regularization methods [22]. In this experiment, one neuron from the hidden layers was dropped. The stratified 10-fold cross-validation was implemented on the same dataset used in the previous experiment, with 189,172 samples as the training set and 21,019 samples as the test set in each fold. Fig. 7 shows the results after applying the dropout technique on the hidden layers. The mean accuracy of 90% obtained reflects an encouraging improvement in building the ADNN.

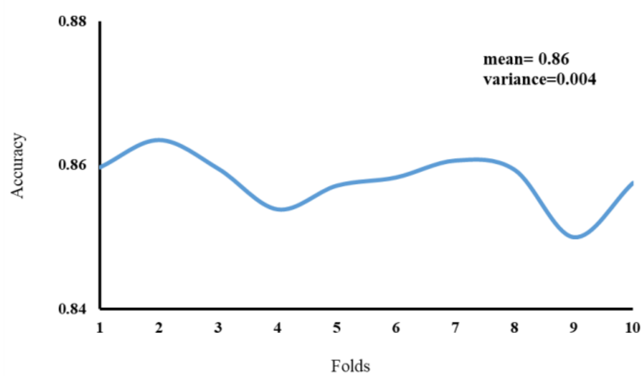


Fig. 5. 10-Fold Cross-Validation Accuracies for Fitting the Training Set.

	Positive	Negative
Positive	TP=20,230	FN=7066
Negative	FP=950	TN=21,754

Fig. 6. Confusion Matrix for Classifying the Test Set in the Second Experiment.

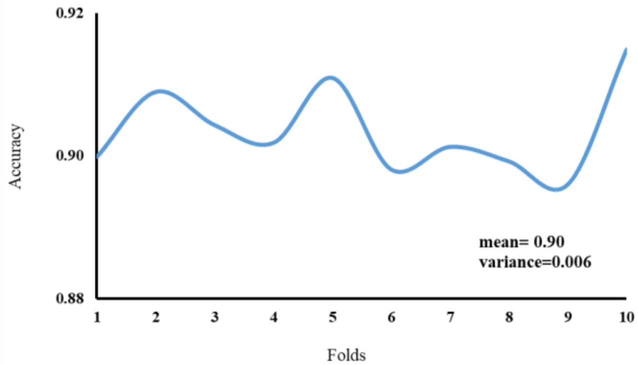


Fig. 7. 10-Fold Cross-Validation Accuracies for Fitting the Training Set after Applying the Dropout Technique.

To verify the ADNN performance on unseen samples, a test was conducted using the same test set used in the previous experiment (50,000 samples). Figure 8 shows the confusion matrix resulting from the test. The accuracy of 90% obtained confirms the enhancement in the ADNN after applying the dropout technique.

D. Parameter Tuning

Despite the improvements achieved in building up the ADNN, there was still room to enhance the prediction accuracy. The best tool to use to achieve a higher accuracy than 90% at this stage was parameter tuning. The ADNN has two types of parameters: 1) tweaking parameters, i.e. the weights learned from the model during the training and 2) hyperparameters, i.e. number of epochs, batch size, the optimizer, and the number of neurons in the layers.

The technique of grid-search cross-validation (GSCV) [2] was used to find the optimal hyperparameters of a neural network that result in the most accurate prediction. The GSCV technique tests several combinations of hyperparameters values and returns the best selection choice that leads to the best accuracy. The GSCV technique usually takes a long time to test the values and can be computationally expensive in case of huge dataset and the number of hyperparameters to be tuned is large. To avoid this, the training phase involved only three hyperparameters, the number of epochs, the batch size, and the optimizer. Table III describes the hyperparameters and the combinations of values that are tested. The number of neurons, number of folds (k), and hidden layers were not changed. The accuracy obtained is 91% for fitting the ADNN to the training set.

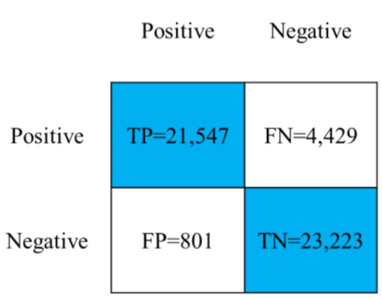


Fig. 8. Confusion Matrix for Classifying the Test Set in the Third Experiment.

TABLE. III. HYPERPARAMETER TUNING

Hyperparameter	Values tested	Best value
Number of epochs	30 and 35	35
Batch size	25 and 32	32
Optimizer	"Adam" and "rmsprop"	"rmsprop"

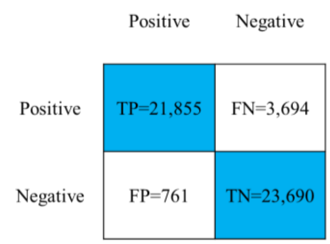


Fig. 9. Confusion Matrix for Classifying the Test Set in the Fourth Experiment.

In the test phase, the same test set (50,000 samples) was used as in the previous experiments, with the accuracy resulting from parameter tuning found to be 91%. Fig 9 shows the confusion matrix resulting from the test.

E. Imbalance Classification Problem

The imbalance classification problem occurs in binary classification when the rate of one class is outnumbered by the other class. Two classes were used in this malicious packet classification, namely normal packets and malicious packets, with the former representing the majority of the dataset. In such a situation, the accuracy is not an optimum measure for assessing the ADNN performance. Two characteristics of the ADNN performance were assessed. First, the ADNN's ability to classify the malicious packets, which are the packets of interest in the dataset. Second, the proportion of packets that the ADNN classifies as malicious that indeed are actually malicious. Recall and precision metrics were used to assess the two characteristics. The recall and precision metrics were identified as follows:

$$recall = \frac{tp}{tp+fn} \quad , \; precision = \frac{tp}{tp+fp}$$

There is a tradeoff between the precision and the recall in binary classification. As the precision increases, the recall decreases and vice-versa. Finding an optimal balance of recall and precision was achieved by combining the two metrics using the F_1 score, which is a harmonic mean of precision and recall that summarizes the model's ability for a specific probability threshold (0.5). The F_1 score was computed as in the following equation.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

The precision-recall curve (PRC) metric was used to summarize the model's performance across more than one threshold. The PRC is a plot of the precision (y-axis) and the recall (x-axis) for different thresholds. Instead of illustrating the curves, the area under the curve (AUC) is calculated. The AUC is an integral summary of the model's performance. A model that performs perfectly has an AUC of 1.0. We compared the accuracy, AUC, recall, precision, and F_1 scores with scores of three commonly used models in the literature: kNN, SVM, and Naïve Bayes. The scores for testing the 52,548 samples with both ADNN and conventional machine learning models are shown Table IV. Fig. 10 shows the TP, FP, FN, and TN scores for ADNN and the three machine learning models. The results show that ADNN is the superior method in terms of accuracy.

TABLE. IV. ACCURACY, AUC, RECALL, PRECISION, AND F1 RESULTS

Algorithm	**Acc**	**AUC**	**Recall**	**Pre**	**F_1**
ADNN	0.92	0.86	0.61	0.96	0.75
kNN (k=5)	0.87	0.83	0.61	0.90	0.73
SVM (kernal='rbf')	0.91	0.85	0.60	0.91	0.72
Naive Bayes	0.89	0.85	0.61	0.86	0.71

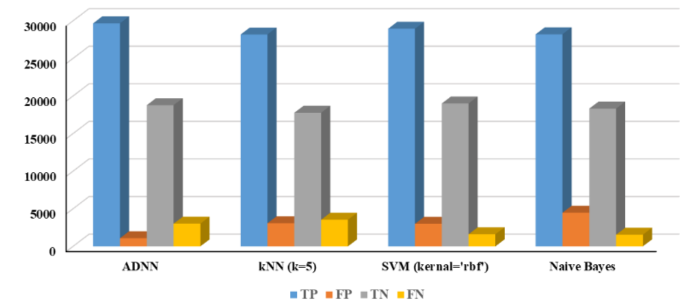


Fig. 10. TP, FP, FN, and TN Results.

V. CONCLUSION

An artificial deep neural network for binary classifying network packets into malicious and normal packets was presented in this paper. The strategy for building up the deep neural network followed systematic stages in order to reach the highest possible level of accuracy. In each stage, an investigation task for a deep learning technique was performed, followed by experiments involving the technique itself. In the final stage, the parameters of the neural network were tuned to confirm the optimum setup. For training and evaluation of the artificial deep neural network, the UNSW-NB dataset was used. The UNSW-NB dataset was created by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security. The preparation of training and test datasets involved four tasks: feature selection, encoding of categorical features, dataset filtration, and feature scaling. The performance was compared with three commonly used models in the literature: kNN, SVM, and Naïve Bayes. The results show that the artificial deep neural network is superior to the competing models in terms of accuracy. Our future research will be directed towards investigating other classes of deep neural networks, e.g. DBN, RNN, and CNN, and applying these algorithms on different public traffic data sets to examine their effectiveness.

REFERENCES

[1] Luo, M., Wang, L., Zhang, H., Chen, J.: A research on intrusion detection based on unsupervised clustering and support vector machine. In: International Conference on Information and Communications Security 2003, pp. 325-336. Springer.

[2] Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press, Cambridge, MA, USA (2016).

[3] lizadeh, H., Zúquete, A.: Traffic classification for managing Applications' networking profiles. Security and Communication Networks 9(14), 2557-2575 (2016).

[4] Naseer, S., Saleem, Y., Khalid, S., Bashir, M.K., Han, J., Iqbal, M.M., Han, K.: Enhanced network anomaly detection based on deep neural networks. IEEE Access 6, 48231-48246 (2018).

[5] Este, A., Gringoli, F., Salgarelli, L.: On-line SVM traffic classification. In: 2011 7th International Wireless Communications and Mobile Computing Conference 2011, pp. 1778-1783. IEEE.

[6] Li, W., Yi, P., Wu, Y., Pan, L., Li, J.: A new intrusion detection system based on KNN classification algorithm in wireless sensor network. Journal of Electrical and Computer Engineering 2014 (2014).

[7] Zhang, J., Xiang, Y., Zhou, W., Wang, Y.: Unsupervised traffic classification using flow statistical properties and IP packet payload. Journal of Computer and System Sciences 79(5), 573-585 (2013).

[8] Nguyen, T.T.T., Armitage, G.: A survey of techniques for internet traffic classification using machine learning. IEEE Communications Surveys & Tutorials 10(4), 56-76 (2008). doi:10.1109/SURV.2008.080406.

[9] Dainotti, A., Pescape, A., Claffy, K.C.: Issues and future directions in traffic classification. IEEE network 26(1), 35-40 (2012).

[10] Shubair, A., Al-Hashmi, A.S.: TiSEFE: Time Series Evolving Fuzzy Engine for Network Traffic Classification. International Journal of Communication Networks and Information Security 10(1), 116-124 (2018).

[11] Donato, W.D., Pescape, A., Dainotti, A.: Traffic identification engine: an open platform for traffic classification. IEEE Network 28(2), 56-64 (2014). doi:10.1109/MNET.2014.6786614.

[12] Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems 2014, pp. 1988-1996.

[13] Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.-Y.: Traffic flow prediction with big data: a deep learning approach. IEEE Transactions on Intelligent Transportation Systems 16(2), 865-873 (2015).

[14] Lotfollahi, M., Zade, R.S.H., Siavoshani, M.J., Saberian, M.: Deep packet: A novel approach for encrypted traffic classification using deep learning. arXiv preprint arXiv:1709.02656 (2017).

[15] Rahul, R., Anjali, T., Menon, V.K., Soman, K.: Deep learning for network flow analysis and malware classification. In: International Symposium on Security in Computing and Communication 2017, pp. 226-235. Springer.

[16] Alom, M.Z., Bontupalli, V., Taha, T.M.: Intrusion detection using deep belief networks. In: 2015 National Aerospace and Electronics Conference (NAECON) 2015, pp. 339-344. IEEE.

[17] Lopez-Martin, M., Carro, B., Sanchez-Esguevillas, A., Lloret, J.: Network Traffic Classifier With Convolutional and Recurrent Neural Networks for Internet of Things. IEEE Access 5, 18042-18050 (2017). doi:10.1109/ACCESS.2017.2747560.

[18] Moustafa, N., Slay, J.: UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: 2015 military communications and information systems conference (MilCIS) 2015, pp. 1-6. IEEE.

[19] Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics 2011, pp. 315-323.

[20] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

[21] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1), 1929-1958 (2014).

[22] Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012).

An Improved Framework for Content-based Spamdexing Detection

Asim Shahzad[1], Hairulnizam Mahdin[*2], Nazri Mohd Nawi[3]
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

***Abstract*—To the modern Search Engines (SEs), one of the biggest threats to be considered is spamdexing. Nowadays spammers are using a wide range of techniques for content generation, they are using content spam to fill the Search Engine Result Pages (SERPs) with low-quality web pages. Generally, spam web pages are insufficient, irrelevant and improper results for users. Many researchers from academia and industry are working on spamdexing to identify the spam web pages. However, so far not even a single universally efficient method is developed for identification of all spam web pages. We believe that for tackling the content spam there must be improved methods. This article is an attempt in that direction, where a framework has been proposed for spam web pages identification. The framework uses Stop words, Keywords Density, Spam Keywords Database, Part of Speech (POS) ratio, and Copied Content algorithms. For conducting the experiments and obtaining threshold values WEBSPAM-UK2006 and WEBSPAM-UK2007 datasets have been used. An excellent and promising F-measure of 77.38% illustrates the effectiveness and applicability of proposed method.**

Keywords—Information retrieval; Web spam detection; content spam; pos ratio; search spam; Keywords stuffing; machine generated content detection

I. Introduction

Spamdexing or web spam is described as " an intentional act that is intended to trigger illegally favorable importance or relevance for some page, considering the web page's true significance" [1]. Studies by different researchers in the area show that on Web at least twenty percent of hosts are spam [2]. Spamdexing is widely recognized as one of the most significant challenges to web SEs [3]. One of the important current problems of SEs is spamdexing because spamdexing heavily reduces the quality of the search engine's results. Many users get annoyed when they search for certain content and ended up with irrelevant content because of web spam. Due to the unprecedented growth of information on the World Wide Web (WWW), the available size of textual data has become very huge to any end user. According to the most recent survey by WorldWideWebSize, the web is consisting of 5.39 billion1 pages. To the web corpus, thousands of pages are being added every day and out of all these web pages several are either spam or duplicate web pages [3]. Web spammers are taking the benefits from internet users by dragging them to their web pages using several smart and creative spamming methods. The purpose of building a spam web page is to deceive the SE in such a way that it delivers those search results which are irrelevant and not beneficial to the web user. The ultimate aim of web spammers is to increase their web page's rank in SERPs. Besides that, spamdexing also has an economic impact because web pages with higher rank can get huge free advertisements and huge web traffic volume. During the past few years, researchers are working hard to develop the new advanced techniques for identification of fraudulent web pages but, spam techniques are evolving also, and web spammers are coming up with new spamming methods every day [4]. Research in the area of web spam detection and prevention has become an arms race to fight an opponent who consistently practices more advanced techniques [4]. If one can recognize and eliminate all spam web pages, then it is possible to build an efficient and robust Information Retrieval System (IRS). Efficient SEs are needed which can produce promising and high-quality results according to the user search query. The next task is to arrange retrieved pages by the content or semantic similarity between retrieved web pages and the search query entered by the user. Finally, the arranged pages are presented to the user. There are many adverse effects of spamdexing on both search engine and end user [5]. Spam web pages not only waste the time but also waste the storage space. As SE needs to store and index a huge number of web pages, so more storage space is required. Furthermore, when SE requires to search web pages on the bases of the user's query, it will search in the huge corpus and therefore more time is required. Due to this, it diminishes the effectiveness of the SE and reduces the trust of the end user on SE [6].

To get over the web spam attacks the improvement in anti-web-spam methods is essential. All the techniques which can be used to get an undeservedly high rank are kind of spamdexing or web spam. Generally, there are three types of spamdexing, Content Spam, Link Spam, and Cloaking. Cloaking is a spamdexing method in which the content offered to the end user is different from that presented to the SE spider [7]. However, the most common types of spamdexing are content and link spam. Content spam is the one studied in this research work. Davison [8] defined the link spam as — the connections among various web pages that are present for a purpose other than merit. In link spam, web spammers are creating the link structure for taking the benefits from link-based ranking algorithms, for instance, PageRank, it will assign the higher rank to a web page if other highly ranked web pages are pointing to the web page with backlinks. Content spam is consisting of all those methods in which web spammers are changing the logical view that an SE has over

[1] https://www.worldwidewebsize.com/
*Corresponding Author

the web page contents [1], it is the most common type of spamdexing [9]. This spamdexing technique is popular among the web spammers because of several SEs are using the information retrieval models (IRM) for instance, statistical language model [10], vector space model [11], BM25 [12] and probabilistic model which are applied to the web page's content for ranking the web page. Spammers try to utilize the vulnerabilities of these models for manipulating the content of target web page [13]. For instance, using important keywords several times on a target web page and increasing the keywords frequencies, copying the content from various good web pages, using the machine-generated content on target web page, and putting all dictionary words on the page and then changing the color of text similar to the background color so users can not see the dictionary words on target page and only visible to SEs spiders are some methods which web spammers are using for getting higher page rank in the SERPs [14]. Generally, content spam can be divided into five subtypes based on the structure of the page. These subtypes are Body spam, Anchor text spam, Meta-tag spam, URL spam, and Title spam. There are a number of spamming methods which are targeting the different algorithms used in SEs [15].

The focus of this article is content spam detection, in this work, a framework has been proposed to detect spam web pages by using content-based techniques. In the proposed approach, stop words, keywords density, Keywords database, part of speech (POS) ratio test, and copied content algorithms are used to detect the spam web pages. For the experimental purpose, WEBSPAM- UK2007 and WEBSPAM-UK2006 datasets are used. The experimental results with an encouraging F-measure demonstrate the applicability and effectiveness of the proposed improved framework as compared to other already existing techniques.

II. Literature Review

In the most recent years, several content-based spamdexing identification methods are proposed by the researchers during the spamdexing challenge [13]. Ntoulas et. al [3] proposed some content features. Their research work showed the text compressibility and HTML based characteristics which identify the content spam from normal web pages. In the research work done by Piskorski et. al [16], they explored a huge number of linguistic features. For text classification tasks the Latent Dirichlet Allocation (LDA) [17] is well known. For spamdexing detection, Biro et al. modified the LDA, they did a lot of research and proposed the linked LDA [18] and multi-corpus LDA [19] models.

For content-based analysis Ntoulas et. al. [3] used the decision tree classifier for identification of spam web pages. Many features were proposed by them for instance, the average length of words, the number of words, anchor text amount and the portion of visible content within in web page. A group of researchers proposed the combinatorial feature fusion and semi-supervised method for identification of spam web pages [20]. For a host, their produced Term Frequency-Inverse Document Frequency (TF-IDF)[2] feature vectors over a hundred pages are all sparse vectors. For exploiting the unlabeled samples, they used the semi-supervised learning, and for creating the new features and reducing the TF-IDF content-based features they used the combinatorial feature fusion method. Empirical results prove the effectiveness of their method.

The most fundamental work on content-based spamdexing detection algorithms is done by Fetterly et al [3], [21]–[23]. In [23] they suggest that using statistical analysis the spam web pages can be classified easily. Because generally, web spammers are generating the spam web pages automatically by adopting the phrase stitching and weaving methods [1], these spam pages are designed for SE spiders instead of real human visitors, so these web pages show the abnormal behavior. Researchers identified that there are a huge number of dashes, dots, and digits in the URL of spam web pages and the length of URL is also exceptional. During their research work, they identify that out of 100 longest observed host-names 80 were pointing to the adult contents, and 11 of them were referring to the financial credit related web pages. They also observe that web pages themselves have the duplicating nature - spam web pages hosted by the same host almost contains the same content with very little variance in word count. Another very fascinating finding of this research was they identified that spammers are changing the content on these spam pages very quickly. Specifically, they kept a close eye on content changing feature and observed the changes on these spam pages for a specific host on weekly bases. They come up with the results that 97.2% of the most active spam hosts can be identified on the bases of this one feature only. They proposed several other features in this research work all other features can be seen in the research article [23].

In another study conducted by them [21], [22], they worked on content duplication and identified that bigger clusters with the identical content are actually spam. For the identification of duplicate content and such clusters they applied the shingling method [24] which is based on Rabin-finger-prints [25], [26]. Initially, they used a primitive polynomial (PA), to fingerprint every *n* words on a web page, secondly, they used a different primitive polynomial (PB), they fingerprint every token from the initial step using extension transformation and prefix deletion, then every string which is obtained from the second step they applied *m* different fingerprint functions on it and hold the tiniest of *n* resulting values for every *m* fingerprint functions. Lastly, the document is a container of *m* fingerprints, so they used transitive closure of the near-duplicate relationship to perform the clustering. In another research study [3], they did some more work and come up with a few more content-based features. Ultimately, they combined all these features in a classification model within bagging frameworks, C4.5 and boosting. For boosting of 10 C4.5 trees they reported 97.8% true negative and 86.2% true positive rates. Another work [27] done by a group of researchers studied the machine learning models and several features, in their research they defined thoroughly that how machine learning models and several features can help in spamdexing identification. They concluded excellent classification results using easy to calculate content features, RandomForest, LogitBoost, and state of the art learning models. They also revealed that global

[2] http://www.tfidf.com/

and computationally demanding features such as PageRank (PR)[3] can help just a bit in quality enhancement. Thus, the researchers claimed that a proper and careful selection of a machine learning model is critical.

To identify the script generated spam web pages Urvoy et. al [28] introduces the features which are based on HTML page structure. They used a very different and non-traditional method for data preprocessing, they removed all the content of the web page and only kept the layout of web page. Therefore, they identified the web page duplication by examining the layout of web page instead of content. Fingerprinting method [25], [26] is applied by them, followed by clustering to identify the groups of spam web pages which are structurally near-duplicate. [29] proposed a method for spamdexing identification in blogs by matching the language models [30] for web pages and blog comments, linked from these comments.

III. Understanding the Content Spam

We believe that it is impossible to tackle the spamdexing without complete knowledge of its working mechanism. The easy target of content-based spamdexing is text relevance algorithms such as TF.IDF [1] and BM25 [12], spammers easily can exploit the weaknesses of these algorithms. These types of algorithms are an easy target for spammers because these algorithms are exposed to content-based spam due to a powerful correlation between document relevance and the number of query words present in the text [31]. Usually, web spammers are using the content-based spamdexing in doorways – websites and web pages which are purposely designed for redirecting and attracting the web traffic [32]. Doorways can only perform effectively if these pages reach the top of SERPs. Normally web spammers like to build hundreds of doorway web pages, and they optimize each doorway web page for a specific keyword to maximize the volume of web traffic collected [31]. To increase the effectiveness of content-based spamdexing there are some requirements[4] which content-based spamdexing must satisfy.

- Content spam pages must be created in hundreds or even thousands. As spammers need to design thousands of spam web pages so these pages can create the content automatically and it will have a lot of spellings error in it [33].
- Each spam web page should increase the text relevancy for a specific search keyword. Therefore, web spammers have a few options for producing the content for their spam web pages. So, spammer can copy the content from other websites [34].
- Each spam page must have designed for generating the profit. Because normally spam web pages are designed for advertisement only with very little and irrelevant content [34].
- Spam web pages do not provide the relevant content to the web users who are browsing these web pages, these spam web pages only targeting the SE spiders [35].
- These web pages must have several irrelevant links and keywords even if these pages are including the real contents [35].
- Must automatically redirect the web users to irrelevant web pages, for instance, a web page which is entirely different from what is expected based on URL, search results or/and anchor text. Sometimes, it can also offer Search Engine Optimization (SEO) services, affiliate links, and random link exchange [36].
- Spam web pages must hide the text using the cloaking techniques, and they should be optimized heavily for search engines.

All web pages can also be categorized as web spam which are only displaying the catalogs of some products but in reality, they are redirecting the users to other traders without giving the additional value. Generating the text automatically is a tough job and there is no satisfactory and good technique available for this yet [37]. There are multiple levels of consistency in natural text, therefore, it is very difficult to follow all at once [38]. Several characteristics of natural texts are distinguished by the researchers during automatic text generation tasks, for instance, automatic document summarization. In different experiments conducted by the researchers, they proved that even very good and highly specialized text generation algorithms [39] didn't perform well and score little in many of these measure [40]. The consistency levels consist of topical consistency, local coherence, logical structure of the document, local coherence etc. In the proposed framework different important components are used for spam identification and tried to make it harder for web spammers to generate the spam web pages.

IV. Data Preprocessing and Experiments for Calculating the Threshold Values

To obtain the suitable threshold values well-known spamdexing datasets WEBSPAM-UK2006 and WEBSPAM-UK2007[5] are used. The WEBSPAM-UK2006 and WEBSPAM-UK2007 collections are based on a crawl of the .uk domain made on May 2006 and May 2007 respectively by the Laboratory of Web Algorithmics[6], Università degli Studi di Milano with the support of the DELIS EU - FET research project. Both datasets are labeled by a group of volunteers. WEBSPAM-UK2006 is consist of 11,402 hosts in total, out of which only 7,473 are labeled. WEBSPAM-UK2007 is consist of 114,529 hosts in total, out of which only 6,479 are labeled. Both datasets divide the web pages into testing and training sets with both non-spam and spam labels. To get optimized threshold values we manually selected 5000 webpages labeled as non-spam/spam by humans. Some content-based features for example Stop words, Keywords, Spam Keywords, Part of Speech and Duplicate Content were extracted for content-based spamdexing identifications. Finally, through different experiments, the most appropriate threshold values were obtained that provide the fewest false positive ratios and high F-measure.

[3] https://www.geeksforgeeks.org/page-rank-algorithm-implementation/
[4] http://chato.cl/webspam/datasets/uk2007/guidelines/
[5] http://chato.cl/webspam/datasets/
[6] http://law.di.unimi.it/

A. Stop words Threshold Value

Commonly used words (for instance, "a", "the", "in", "that", "an") are known as Stop words. Usually, programmers programmed search engines to ignore Stop words during indexing the records for searching. Stop words are considered irrelevant for searching purpose because Stop words frequently occur in natural language. To save time and space Stop words are dropped during indexing and ignored at search time. Spammers are taking advantage of this, to get a higher rank on Search Engine Result Pages (SERPs) they are generating the machine-generated articles which contain a high frequency of repeated keywords with less or no Stop words in the article. Using Stop words spamdexing can be detected. We are considering this feature for content-based spamdexing detection. To calculate the threshold value, approximately three thousand human labeled non-spam web pages were selected manually. A script has been used for Stop words identification and counting on each non-spam web page chosen for this experiment. The ratio of stop words on each web page is calculated using equation (1).

$$RSW\ on\ Wpi = \frac{No\ of\ SW\ on\ Wp_i}{Total\ number\ of\ words\ in\ Wp_i} \quad (1)$$

Where RSW represents the Ratio of Stop words, Wp_i is any webpage, and SW represents the Stop words. Finally, the average ratio of all Stop words is calculated and used as a standard threshold value for Stop words. Average Ratio = Sum of ratios of all Stop words / Total number of pages.

$$Stopwords\ Threshold\ Value = \frac{RSW_1 + RSW_2 + \cdots + RSW_i}{Wp_1 + Wp_2 + \cdots + Wp_i} \quad (2)$$

B. Keywords Threshold Value

Keyword frequency represents the percentage of times a phrase or Keyword appears on a webpage compared to the total number of words on the webpage. In search engine optimization context, keyword frequency can be used to check if a web page is relevant to a specific keyword or phrase. Some of the search engine optimization experts are saying that the optimal keywords density is unknown until Google or any other big search engine reveals it. And several search engine optimization experts think that the best keyword frequency is one to three percent and more could be regarded as search spam. After reading different research articles, search engine's guidelines and consulting the SEO experts, we found a lot of different conflicting opinions on ideal keyword frequency ratio. So we decided to calculate our threshold value for keywords frequency ratio through experiments. Around three thousand humans labeled non-spam web pages were manually selected from the training data set for calculating the threshold value. A script has been used for identification and storing of Keywords from every webpage Wp_i on a file. After Keywords identification, Keywords frequency test is performed. This test is used to determine the frequency (in percentage) of each distinct Keyword KW on a webpage Wp_i, i.e., the percentage of occurrences of every distinct Keyword compared to all other keywords on a webpage Wp_i. The steps given below shows how Keywords frequency test is performed on a webpage Wp_i.

1) Identify the Keyword frequency of every distinct keyword on a webpage Wp_i. The Keyword frequency of a distinct Keyword KW_i on webpage Wp_i is defined as:

$$T_{kw} = KW_1 + KW_2 + KW_3 + \cdots KW_i$$

$$KW_f(KW_i) = \frac{No\ of\ times\ KW_i\ appears\ on\ Wp_i}{T_{kw}\ on\ Wp_i} \quad (3)$$

Whereas KW_i represents the Keyword, Wp_i represents Webpage, T_{kw} is total number of keywords on Wp_i, and KW_f is keyword frequency.

2) Identify the Keyword Phrase frequency of every Keyword Phrase on a webpage Wp_i. The Keyword Phrase frequency of a distinct Keyword Phrase KW_p on webpage Wp_i is defined as:

$$KW_f(KW_p) = \frac{No\ of\ times\ KW_i\ appears\ on\ Wp_i * N_{wp}}{T_{kw}\ on\ Wp_i} \quad (4)$$

Whereas KW_p represents the Keyword Phrase, and N_{wp} is the number of words in a phrase.

3) Finally, after obtaining the frequency of every Keyword on all webpages selected for the experiment, the average frequency of all keywords is calculated to get a standard threshold value.

$$Keywords\ Threshold\ Value = \frac{KW_f(KW_1) + KW_f(KW_2) + \cdots + KW_f(KW_i)}{Total\ number\ of\ KW_f} \quad (5)$$

C. Spam-Keywords Threshold Value

Spam Keywords can play a significant role in spamdexing detection and prevention. Web spammers are excessively using Spam Keywords on their websites, pages, newsletters, and emails. Spam Keywords can be categorized in different categories such as personal, general, financial, medical, free offers, sense of urgency, exaggerated claims, etc. The Spam Keywords are used for content-based spamdexing detection in the proposed improved framework. The steps given below shows how the threshold value for Spam Keywords is calculated.

1) The initial step was to create a database of Spam Keywords. Several researchers, SEO experts, and search engines already identified the Spam Keywords spammers are commonly using in their content. After consulting different experts in the field and checking the existing well-known Spam Keywords Databases[7, 8, 9] , a more significant Spam Keywords database was created.

2) For calculating the threshold value, we manually selected approximately three thousand human labeled non-spam web pages from the dataset. A script has been used to identify all Spam Keywords using Spam Keywords database on all webpages selected for calculating the threshold value.

[7] https://www.automational.com/spam-trigger-words-to-avoid/

[8] https://blog.hubspot.com/blog/tabid/6307/bid/30684/the-ultimate-list-of-email-spam-trigger-words.aspx

[9] https://help.emarsys.com/hc/en-us/articles/115005000225-Known-spam-keywords

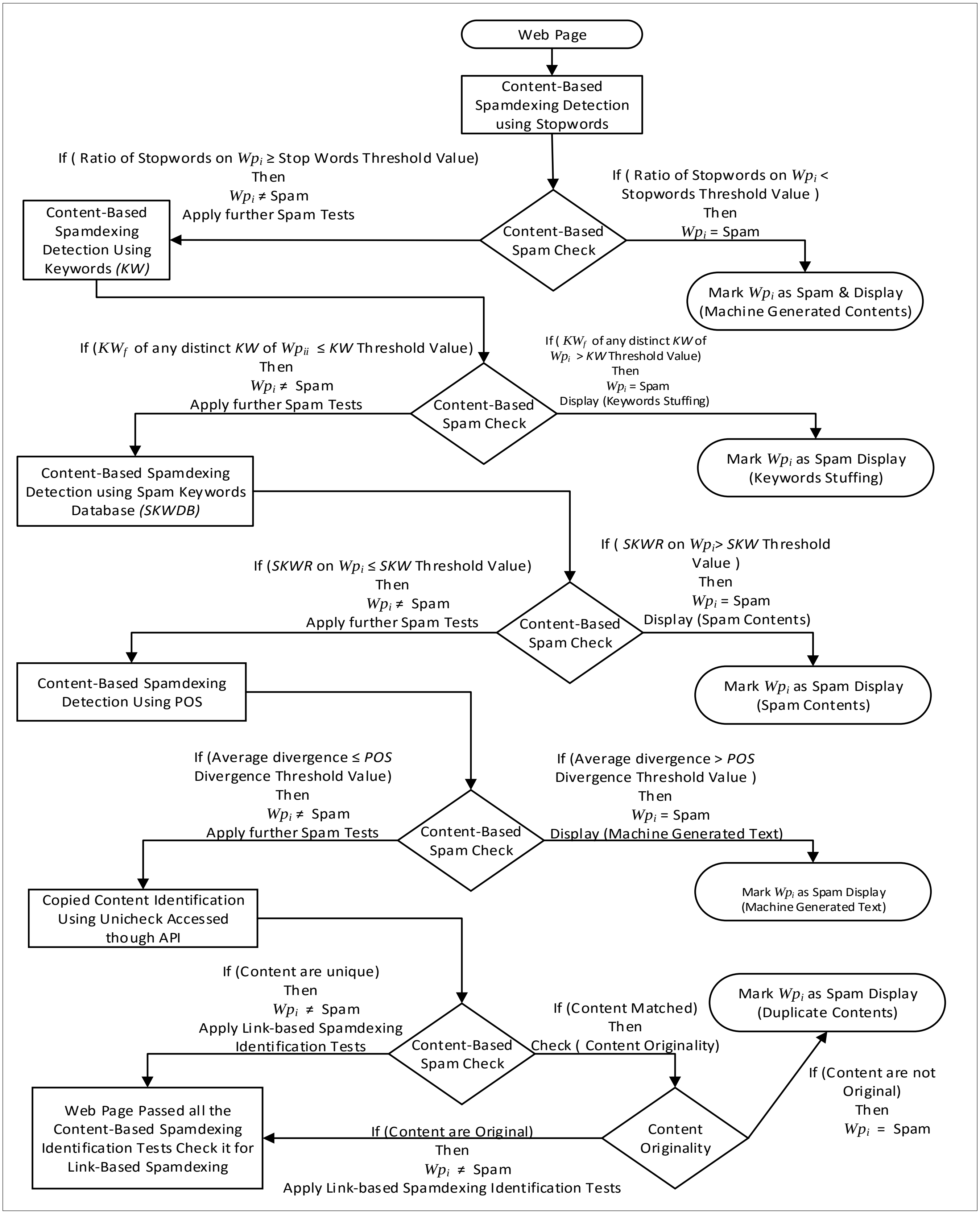


Fig. 1. Improved Framework for Content-based Spamdexing Detections.

3) After Spam Keywords identification, the ratio of Spam Keywords is calculated on each page using equation (6).

$$SKWR = \frac{T_{SKW}}{T_{kw} \text{ on } Wp_i} \tag{6}$$

Where *SKWR* represents the Spam Keywords Ratio, and T_{SKW} represents the total number of spam keywords on webpage Wp_i.

4) Finally, after obtaining the Spam Keywords Ratio on all webpages selected for the experiment, the average Spam Keywords Ratio was calculated to get a standard threshold value.

$$Spam\ Keywords\ Threshold\ Value = \frac{SKWR_1 + SKWR_2 + \cdots + SKWR_i}{Total\ number\ of\ SKWR} \tag{7}$$

D. Part of Speech (POS) Threshold Value

Words are the smallest elements in the natural language that have unique meanings. These words can be categorized into various types based on their functions and use. In the English language, there are eight significant parts of speech: adjective, verb, adverb, interjection, conjunction, preposition, noun and pronoun. There is a standard ratio for each part of speech in the English Language. A researcher Eric Lease Morgan performed the experiments and calculated the standard ratio of occurrence of each grammatical form in English. For the standard ratio figures, the website[10] has been referred.

E. Copied Content

Copied content appears in more than one place on the Internet. For search engines, it is difficult to decide which version of the content is more suitable to the search query. To get high rank on SERPs usually, web spammers are copying the content from other good websites and using that content on their spam pages without or very little change in the content. A custom script has been used for the identification of copied content on the Internet. This script is using Unicheck API[11], which allowed to integrate Unicheck into the content workflow for instant identification of originality of the content as it enters the proposed system.

V. Content-Based Spamdexing Detection Framework

In this section, an improved framework for content-based spamdexing detection is presented. Figure 1 is showing an overview of the proposed framework. This framework is using five different methods, and every method is using a unique feature for identification of content-based spamdexing. The methods are as follows:

1) Content-Based Spamdexing Detection using Stop words.

2) Content-Based Spamdexing Detection using Keywords.

3) Content-Based Spamdexing Detection using Spam Keywords Database (SKWDB).

4) Content-Based Spamdexing Detection Using POS.

5) Copied Content Identification Using Unicheck.

All five techniques are discussed in detail in the following subsections.

A. Spamdexing Detection using Stop Words

The first method is using stop words feature for identification of machine-generated text. A custom script has been used, which is accepting web page content as input and generating two separate output files (stop words and keywords file). Table 1 is showing an example of how this script works.

After obtaining the stop words file, it will count the number of stop words in the file for calculating the ratio of stop words on Wp_i using equation (1). Based on the stop words ratio on Wp_i, the Wp_i will classify into one of the two different categories specified below.

1) Category 1 (Low Stop Words Ratio): For generating the content quickly, web spammers are using the machine-generated content. After keywords research using different tools, they are producing the machine-generated articles [41] with no or very little, randomly inserted stop words in these articles. If the ratio of stop words on webpage Wp_i is lower than stop words threshold value (determined by experiment in sub-section 4.1) then web page $Wp_i \in$ *spamdexing*.

By using the stop words ratio, the proposed method can efficiently identify the machine-generated web pages. Figure 2 is an example of this type of spam web page.

TABLE. I. The Example of the Working Mechanism of Custom Script

Input	*Output*	
Web Page Content	*Keywords file*	*Stop words file*
UTHM is the best University in Malaysia.	UTHM, best, University, Malaysia	is, the, in
It is the public sector University in Batu Pahat.	public, sector, University, Batu Pahat	it, is, the, in
The motto of UTHM is We produce professionals.	motto, UTHM, produce, professionals	the, of, is, we

Fig. 2. Machine-Generated Webpage without Stop Words.

[10] http://infomotions.com/blog/2011/02/forays-into-parts-of-speech/
[11] https://unicheck.com/plagiarism-api-documentation

2) Category 2 (Equal or High): If the ratio of stop words on webpage Wpi is equal or higher than stop words threshold value (determined by experiment in sub-section 4.1) then web page Wpi passed the initial content-based spam test, and it will be submitted to the next web spam detection method for further tests. Figure 3 represents the algorithm for the content-based spamdexing detection using stop words.

B. Spamdexing Detection using Keywords

Keywords frequency test is the method to detect the keywords stuffing. This method is used to determine the frequency (in %) of every distinct keyword on webpage Wp_i. Keywords frequency test is done on webpage Wp_i using equations (3) and (4). Based on the keyword's frequency of every distinct keyword $KW \in Wp_i$, the web page Wp_i will classify into one of the two separate categories discussed below.

1) Category 1 (High Keyword Frequency): If the keyword frequency of any distinct keyword of web page Wp_i is higher than the keywords threshold value (determined by experiment in sub-section 4.2) then web page $Wp_i \in spamdexing$. All web pages belong to this category should be eliminated directly because it clearly shows that keyword stuffing is used which is content-based spamdexing technique. Keyword stuffing means, the ratio of different keywords in Wp_i is very less and the web spammer has repeatedly used some of the keywords too many times in Wp_i [42]. Keyword stuffing is done by repeating some specific distinct keyword/keywords again and again in several spots of Wp_i for instance in Alt attributes, content, comment tags, and Meta tags [42]. The primary target of a web spammer is to make search engine consider that the web page is relevant to different keywords present in the user's query and thus improve its page ranking. Therefore, Wp_i should be categorized as spam. Figure 4 is an example of keywords stuffing.

2) Category 2 (Low or Equal to Threshold Value): If the keyword frequencies of all different keywords of a web page Wp_i are less than KW threshold value (determined by experiment in sub-section 4.2) then web page Wp_i passed the second content-based spam test, and it will be submitted to the next web spam detection method for further tests. Figure 5 represents the algorithm for content-based spamdexing detection using keywords.

C. Spamdexing Detection using Spam Keywords Database

Content-based spamdexing can be detected using spam keywords database. A custom script has been used for identification of spam keywords on web page Wp_i. The script is matching every distinct spam keyword from spam keywords database with every distinct keyword in Wp_i for marking the spam keywords on Wp_i. After marking spam keywords, it calculates the spam keyword ratio *(SKWR)* on Wp_i using equation (6). Based on *SKWR* on Wp_i, the Wpi will classify into one of the two different categories specified.

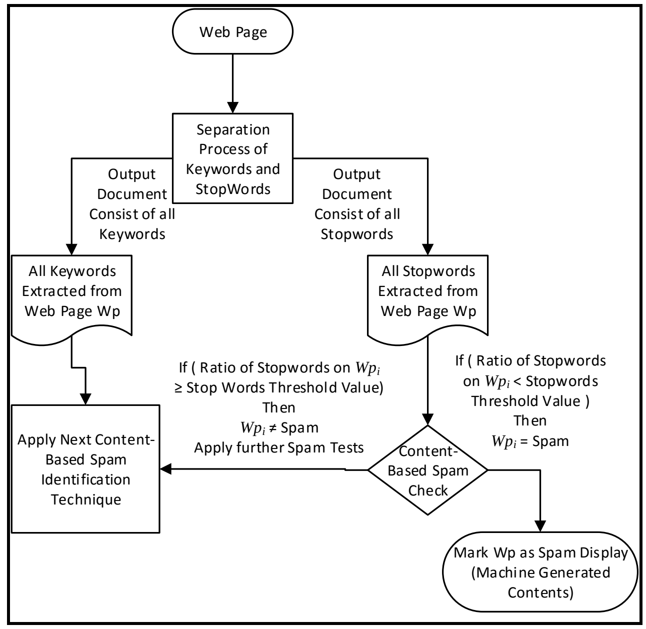


Fig. 3. Algorithm for Content-based Spamdexing Detection using Stop Words.

Fig. 4. Keyword Stuffing Example.

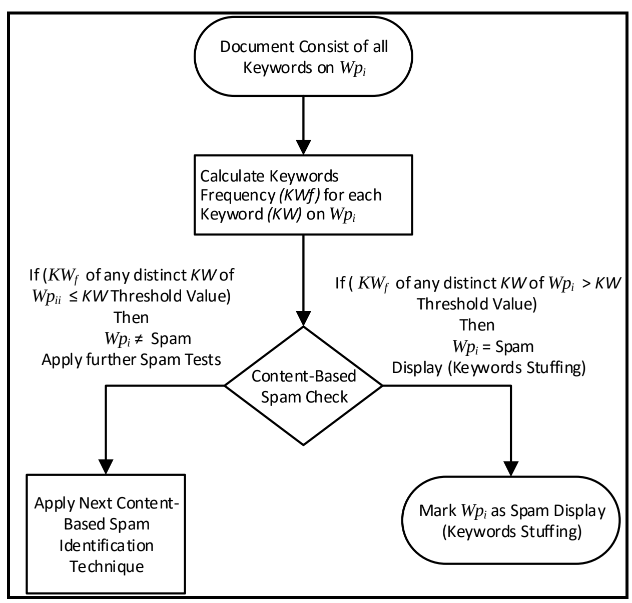


Fig. 5. Algorithm for Content-based Spamdexing Detection using Keyword Frequencies.

1) Category 1 (High Spam Keywords Ratio): If *SKWR* on Wp_i is higher than *SKW* threshold value (determined by experiment in sub-section 4.3) then web page $Wpi \in spamdexing$. All pages belong to category 1 will be eliminated directly and marked as spam because it clearly shows the high usage of spam keywords which is a type of content-based spamdexing. Figure 6 is an example of spam keywords.

2) Category 2 (Low or Equal to Threshold Value): If *SKWR* on *Wpi* is less than or equal to *SKW* threshold value (determined by experiment in sub-section 4.3) then web page *Wpi* passed the third content-based spam test, and it will be submitted to the next web spam detection method. Figure 7 represents the algorithm for content-based spamdexing detection using spam keywords.

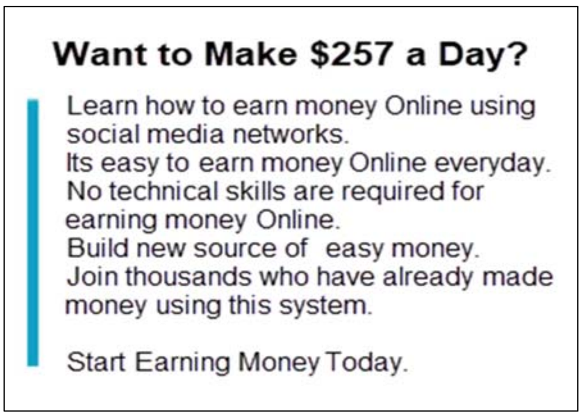


Fig. 6. Spam Keywords usage Example.

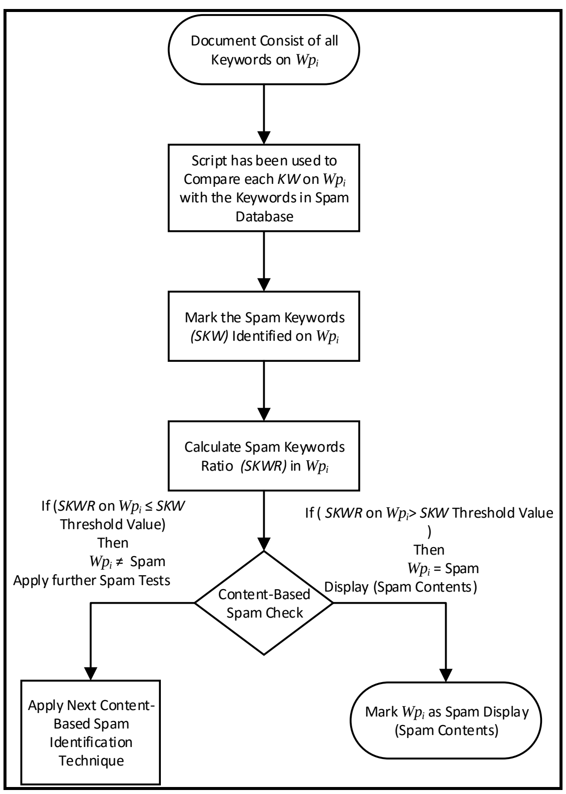


Fig. 7. Algorithm for Content-based Spamdexing Detection using SKWDB.

D. Spamdexing Detection using Part of Speech (POS)

Content repurposing can be detected using a linguistic technique known as part of speech (POS) [43]. Nowadays web spammers are using very advanced techniques for a content generation; machine-generated content is one of them. Some of the spam web pages on the internet are machine generated. Spammers are creating these spam web pages by combining large portions of a single page or by including various small sections of a page into a single web page. To detect the content repurposing on the web page linguistic features can be applied. This technique depends on the supposition that web spammers cannot replicate every aspect of natural language while producing machine generated content. The significance of utilizing broad ranges of linguistic features are discussed by Piskorski et al. [16] in their work. The primary purpose of applying these linguistic features is to recognize the originality and authorship of the content of a page. There are various grammatical forms (g_f) for instance, adjective, adverb, verb, pronoun, noun, conjunction, preposition, and interjection [44]. The ratio of every grammatical form is calculated to obtain the maximum information. The part of the speech ratio test is performed on web page *Wpi* as follows:

1) Finding and tagging the various g_f in Wpi. To tag each word in *Wpi*, Stanford Log-linear Part-Of-Speech Tagger [45] has been used.

2) Grammatical form ratio calculation: The ratio of every grammatical form, $g_f \in Wpi$ is calculated as follows:

$$Ratio\ (g_f) = \frac{x}{y} \tag{8}$$

Where x represents the number of occurrences of g_f in *Wpi* and *y* is the total number of words present in *Wpi*.

3) Calculation of divergence: initially the divergence of the ratio of g_f from the standard ratio of the existence of g_f available in standard English text is calculated. The standard value of each g_f can be seen on this website[10].

4) Calculation of average divergence: After computing the divergence of every $g_f \in Wpi$, the average divergence of *Wpi* is calculated using equation (9).

$$Average\ divergance = \frac{a}{b} \tag{9}$$

where *a* represents the sum of divergence of every $g_f \in Wpi$ and *b* is the total number of grammatical forms considered on *Wpi*.

5) Checking the spam status: Finally, it is time to perform the POS test. Based on the average divergence of *Wpi*, the *Wpi* will classify into one of the two different categories specified below.

a) Category 1 (average divergence is grater or equal to threshold value): If the average divergence is higher or equal to POS divergence threshold value (determined by experiments in sub-section 4.4) then web page Wpi ∈ spamdexing, all web pages belong to this category will be eliminated directly and marked as spam because *Wpi* fails to qualify the part of the speech ratio test.

b) Category 2 (Average divergence is less than threshold value): If average divergence is less than POS divergence threshold value (determined by experiment in sub-section 4.4) then *Wpi* passed the fourth content-based spam test, and it will be submitted to the next content-based web spam detection technique. Figure 8 represents the algorithm for content-based spamdexing detection using POS.

E. Spamdexing Detection using Copied Content

Usually, the web spammers are copying the content from other similar pages and using copied content on their spam web pages. Content-based spamdexing can be identified through copied content. The copied content test is performed on web page *Wpi* is as follows:

1) A custom script has been used to access Unicheck for duplicate content identification. This integration with Unicheck is done through API.

2) The custom script accepts URL of a page as input and identifies all the pages on the internet with similar content. It returns the URLs of all pages having the duplicate content on it.

3) Based on the content of *Wpi*, the *Wpi* will classify into one of the two different categories specified.

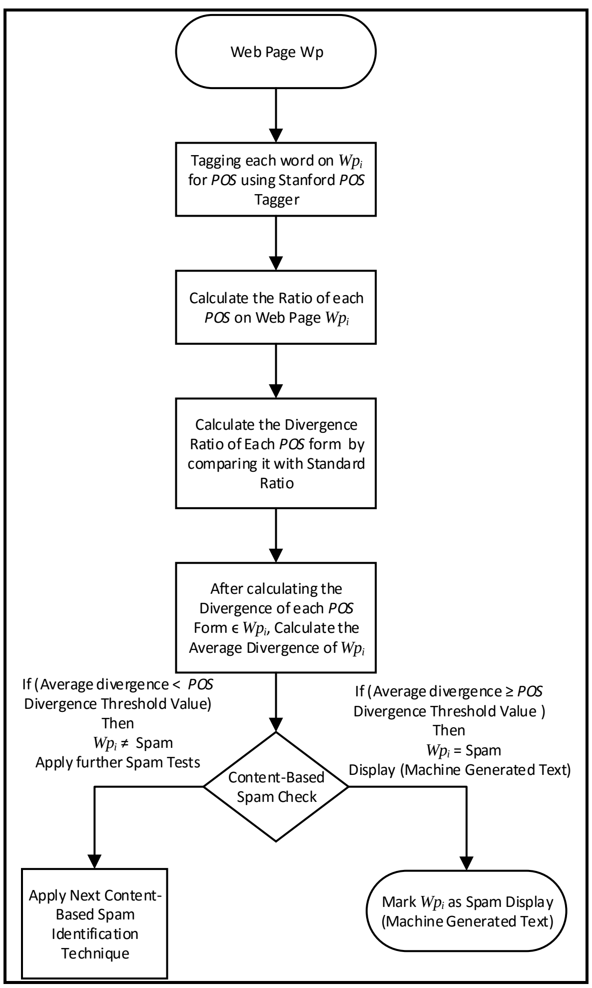


Fig. 8. Algorithm for Content-based Spamdexing Detection using POS.

a) Category 1 (Original Content): If the content of web page *Wpi* is unique and original, then *Wpi* passed all the content-based spamdexing tests, and *Wpi* is not a spam web page. It will be submitted to the link-based spam detection techniques for further testing. To save time, we first performed content-based spamdexing detections tests to identify the spam web pages (if any web page *Wpi* fails in any of the five tests then it is declared to be web spamming, and no further analysis is required). Initially applying the content-based identification techniques reduces the number of web pages for link-based tests.

b) Category 2 (Duplicate Content): If the content of web page *Wpi* is not unique, then check the originality and authorship of the content. For checking the originality and authorship, the publishing date is essential. A web page *Wpi* will be considered original if it is published before all the other duplicate web pages. Usually, the publishing date of a web page is available on the bottom or top of the same page. So, it can be easily identified that when the page was published. But there are some pages on the Internet without a publishing date on it, for finding the publishing date of such web pages we implemented a custom script using a small Google hack. Every web page published publicly on the Internet is having three different dates (publication, indexed and cache date) associated with it. To find the publication date of a web page our script is working as follows:

1) Open https://www.google.com and will paste the URL of web page *Wpi* in the search box with operator inurl: e.g. inurl:www.uthm.my/contact-us. Click search icon for searching.

2) After getting the search results for the URL above, go to the browser's address bar and at the end of Google search URL paste &as_qdr=y15 and click the search icon again for searching.

3) Google search engine will load the SERP again, and this time it will show the actual publication date of *Wpi* next to the title. Figure 9 shows the publication date of the URL.

After performing the originality and authorship test if the content is original then web page *Wpi* is not a spam page and it will be submitted to link-based techniques for further testing, and if the material is copied or not unique, then the web page is classified as spam, and no further testing are required. Figure 10 shows the algorithm for content-based spamdexing detection using copied content.

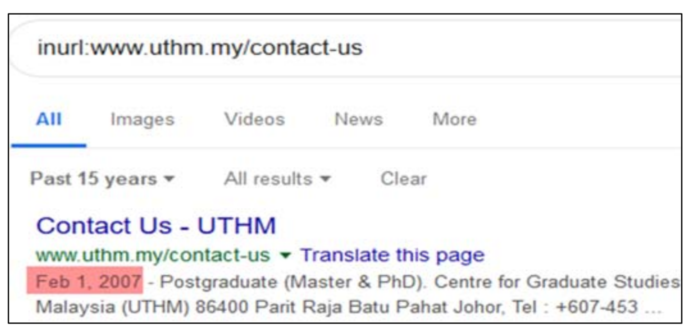


Fig. 9. Example of Publication Date.

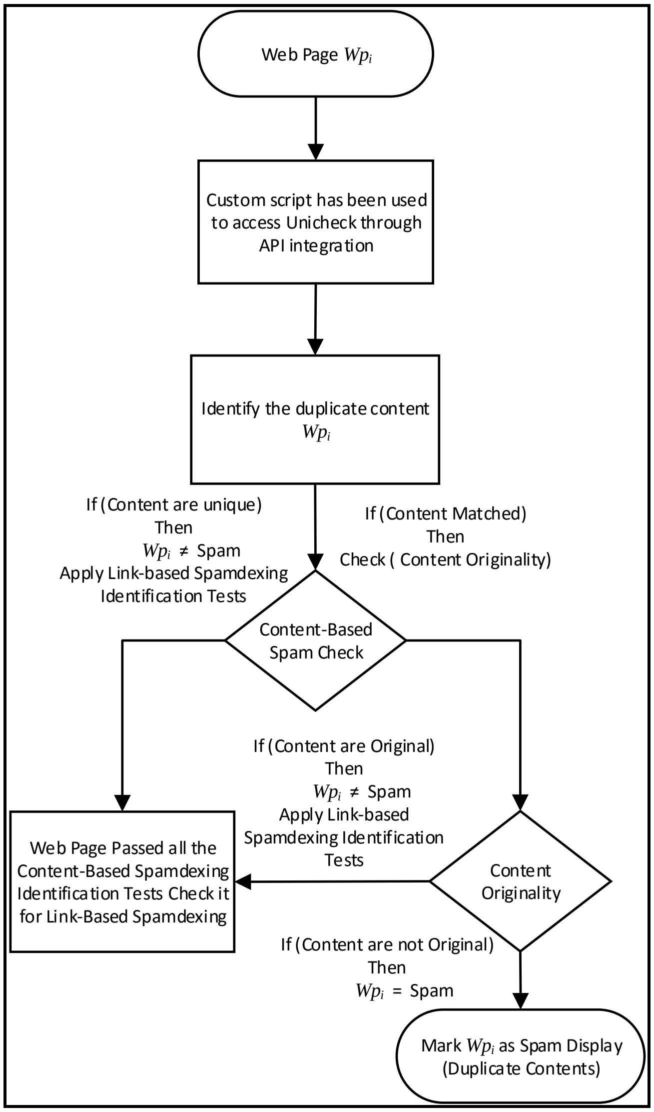


Fig. 10. Algorithm for Content-based Spamdexing Detection using Copied Content.

VI. Experimental Results

For conducting the experiments, the verification set is consisting of randomly chosen web pages which are labeled as non-spam and spam. These web pages are selected from dataset WEBSPAM-UK2006 and WEBSPAM-UK2007. These datasets are well-known and perfectly suited for web spam detection because of the following properties:

1) The datasets are consisting of several types of non-spam and spam web pages.

2) The dataset is freely available to all the researchers in the field and is used as a benchmark measure in the identification of spam web pages.

3) In the datasets, the sample web pages are uniform and random.

4) It includes different types of spam web pages which are produced by using several types of web spamdexing methods.

5) The datasets split the pages into testing and training sets with both non-spam and spam labels so these datasets can effectively be used for any link or content-based technique.

6) To get the optimized threshold values used for the proposed improved framework for content-based spamdexing detection these datasets has been utilized.

WEBSPAM-UK2006 is consist of 11,402 hosts in total, out of which only 7,473 are labeled. WEBSPAM-UK2007 is consist of 114,529 hosts in total, out of which only 6,479 are labeled. By practicing the below pre-processing methods, we obtained the dataset of five thousand pages.

1) Only those pages are considered which are labeled as non-spam or spam by real humans.

2) Among the human labeled pages, we only selected those web pages which are currently existing/working links.

3) We further filtered out the web pages and only selected those pages which are having at least 1KB content, which is necessary for our improved framework for content-based spamdexing detection.

4) Finally, we extracted the content of these web pages and stored the content in text file format.

Python is used for implementing the proposed framework, and a machine with 2x Intel Xeon E5-2670 V2 2.5GHz 10 Core, with 128GB DDR3 and operating system Ubuntu 14.04 has been used for the execution of algorithms. As F-measure is a standard approach for combining both precision and recall, so for comparison of the proposed work with other similar related works and for evaluation of the proposed algorithm we used the F-measure. The proposed improved framework for content-based spamdexing detection achieved the results shown in table 2.

TABLE. II. Performance Evaluation of Improved Framework

Technique	Precision (%)	Recall (%)	F-Measure (%)
The Proposed Framework for Content-Based Spamdexing Detection	78.3	75.6	77.4

VII. Comparison with Existing Approaches

The experimental results of the proposed improved content-based framework are compared with the following existing approaches. The comparison in Table 3 clearly shows that the proposed framework outperforms other spam detection methods. Figure 11 shows the comparison of all techniques.

1) Our proposed framework vs Roul et al [15]: The results of the proposed framework is compared with the research work of Roul et al [15]. For detecting the content-based spam web pages, they have used two features (keywords and POS). As per table 2 of Roul et al [15] they achieved an F-measure of about 70.2% and precision of approximately 71.3%, which is significantly less than our results.

TABLE. III. COMPARISON OF THE PROPOSED FRAMEWORK WITH OTHER STANDARD TECHNIQUES

Content-based Spam detection techniques	Precision	Recall	F-measure
The Proposed Improved Framework	78.3	75.6	77.38
Roul et al	71.3	69.3	70.2
Dai et al	65	44.3	52.7

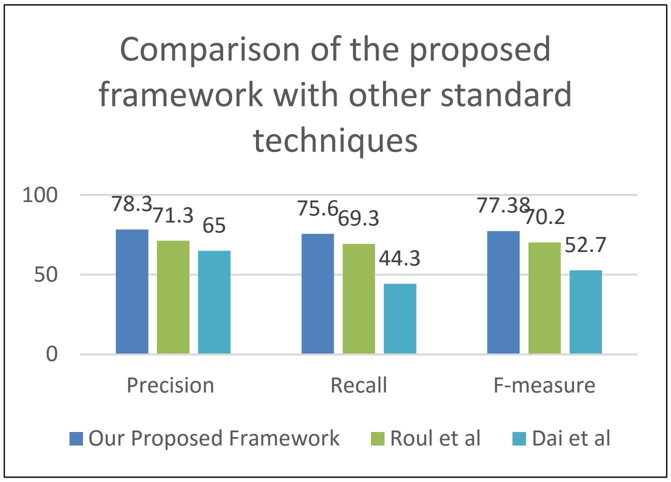


Fig. 11. Comparison of the Proposed Framework with other Standard Techniques.

2) Our proposed framework vs Dai et al [46]: Next, we compared our empirical results with Dia et al [46]. For spam identification they considered the historical web page information in their work. For improvement in spam classification, they have used the content features from the old version of pages. By applying the supervised learning techniques, they combined the classifiers based on the temporal characteristics and the current page content. With their method, they extracted several temporal features from archival copies of the web presented by Internet Archive's Way Back Machine. For their experiments, they have used the dataset WEBSPAM-UK2007. As per table 3 of Dai et al [46] they achieved an F-measure of about 52.7 and precision of approximately 65, which is less than our results.

VIII. CONCLUSION AND FUTURE WORK

In the lives of Internet users, the spamdexing is becoming a very big issue and causing big financial losses. Several techniques have been proposed to detect the spamdexing automatically and avoid this issue. In this article, we proposed an improved framework for content-based web spam detection. We explored five different techniques namely, stop words density, keywords density, spam keywords density, part of the speech ratio and copied content test to detect a web page as non-spam or spam. For this experimental work, we have used two datasets WEBSPAM-UK2006 and WEBSPAM-UK2007. An excellent and very promising F-measure of 77.38% compared to other existing approaches shows the robustness of our framework. We will extend this research work by adding the link-based spamdexing detection techniques to this framework. We believe that by using combine technique we can enhance the power of our framework to identify the wide range of web spam pages.

ACKNOWLEDGMENT

The authors would like to thank Ministry of Higher Education (MOHE) Malaysia and Universiti Tun Hussein Onn Malaysia (UTHM) for financially supporting this Research under Fundamental Research Grant (FRGS) vote 1611.

REFERENCES

[1] Z. Gyongyi and H. Garcia-Molina, “Web spam taxonomy,” in First international workshop on adversarial information retrieval on the web (AIRWeb), 2005.

[2] M. R. Henzinger, R. Motwani, and C. Silverstein, “Challenges in web search engines,” in IJCAI, 2003, vol. 3, pp. 1573–1579.

[3] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, “Detecting spam web pages through content analysis,” Proc. 15th Int. Conf. World Wide Web - WWW ’06, p. 83, 2006.

[4] N. Z. J. MCA and P. Prakash, “Document content based web spam detection using cosine similarity measure,” vol. 7, no. June, 2016.

[5] A. Shahzad, N. M. Nawi, E. Sutoyo, M. Naeem, A. Ullah, S. Naqeeb, and M. Aamir, “Search Engine Optimization Techniques for Malaysian University Websites: A Comparative Analysis on Google and Bing Search Engine,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 4, pp. 1262–1269, 2018.

[6] Y. Li, X. Nie, and R. Huang, “Web spam classification method based on deep belief networks,” Expert Syst. Appl., vol. 96, pp. 261–270, 2018.

[7] Z. Guo and Y. Guan, “Active Probing-Based Schemes and Data Analytics for Investigating Malicious Fast-Flux Web-Cloaking Based Domains,” in 2018 27th International Conference on Computer Communication and Networks (ICCCN), 2018, pp. 1–9.

[8] B. Davison, “Recognizing nepotistic links on the web,” Artif. Intell. Web Search, pp. 23–28, 2000.

[9] N. Spirin and J. Han, “Survey on web spam detection: principles and algorithms,” Acm Sigkdd Explor. Newsl., vol. 13, no. 2, pp. 50–64, 2012.

[10] C. Zhai, “Statistical language models for information retrieval,” Synth. Lect. Hum. Lang. Technol., vol. 1, no. 1, pp. 1–141, 2008.

[11] G. Salton, A. Wong, and C.-S. Yang, “A vector space model for automatic indexing,” Commun. ACM, vol. 18, no. 11, pp. 613–620, 1975.

[12] S. Robertson, H. Zaragoza, and M. Taylor, “Simple BM25 extension to multiple weighted fields,” in Proceedings of the thirteenth ACM international conference on Information and knowledge management, 2004, pp. 42–49.

[13] N. El-Mawass and S. Alaboodi, “Data Quality Challenges in Social Spam Research,” J. Data Inf. Qual., vol. 9, no. 1, pp. 4:1--4:4, 2017.

[14] A. Shahzad, N. M. Nawi, N. A. Hamid, S. N. Khan, M. Aamir, A. Ullah, and S. Abdullah, “The Impact of Search Engine Optimization on The Visibility of Research Paper and Citations,” JOIV Int. J. Informatics Vis., vol. 1, no. 4–2, pp. 195–198, 2017.

[15] R. K. Roul, S. R. Asthana, and M. I. T. Shah, “Detection of spam web page using content and link-based techniques : A combined approach,” vol. 41, no. 2, pp. 193–202, 2016.

[16] J. Piskorski, M. Sydow, and D. Weiss, “Exploring linguistic features for web spam detection: a preliminary study,” in Proceedings of the 4th international workshop on Adversarial information retrieval on the web, 2008, pp. 25–28.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” J. Mach. Learn. Res., vol. 3, no. Jan, pp. 993–1022, 2003.

[18] I. Bíró, D. Siklósi, J. Szabó, and A. A. Benczúr, “Linked latent dirichlet allocation in web spam filtering,” in Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web, 2009, pp. 37–40.

[19] I. Bíró, J. Szabó, and A. A. Benczúr, "Latent dirichlet allocation in web spam filtering," in Proceedings of the 4th international workshop on Adversarial information retrieval on the web, 2008, pp. 29–32.

[20] Y. Tian, G. M. Weiss, and Q. Ma, "A semi-supervised approach for web spam detection using combinatorial feature-fusion," in Proceedings of the Graph Labelling Workshop and Web Spam Challenge at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery, 2007, pp. 16–23.

[21] D. Fetterly, M. Manasse, and M. Najork, "Detecting phrase-level duplication on the world wide web," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005, pp. 170–177.

[22] D. Fetterly, M. Manasse, and M. Najork, "On the evolution of clusters of near-duplicate web pages," in Web Congress, 2003. Proceedings. First Latin American, 2003, pp. 37–45.

[23] D. Fetterly, M. Manasse, and M. Najork, "Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages," in Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004, 2004, pp. 1–6.

[24] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the web," Comput. Networks ISDN Syst., vol. 29, no. 8, pp. 1157–1166, 1997.

[25] A. Z. Broder, "Some applications of Rabin's fingerprinting method," in Sequences II, Springer, 1993, pp. 143–152.

[26] M. O. Rabin, "Fingerprinting by random polynomials," Tech. Rep., 1981.

[27] M. Erdélyi, A. Garzó, and A. A. Benczúr, "Web spam classification: a few features worth more," in Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality, 2011, pp. 27–34.

[28] T. Urvoy, T. Lavergne, and P. Filoche, "Tracking Web Spam with Hidden Style Similarity.," in AIRWeb, 2006, pp. 25–31.

[29] G. Mishne, D. Carmel, and R. Lempel, "Blocking Blog Spam with Language Model Disagreement.," in AIRWeb, 2005, vol. 5, pp. 1–6.

[30] D. Hiemstra, "Language Models BT - Encyclopedia of Database Systems," L. LIU and M. T. ÖZSU, Eds. Boston, MA: Springer US, 2009, pp. 1591–1594.

[31] A. Pavlov and B. Dobrov, "Detecting content spam on the web through text diversity analysis," CEUR Workshop Proc., vol. 735, pp. 11–18, 2011.

[32] E. S. Swirsky, C. Michaels, S. Stuefen, and M. Halasz, "Hanging the digital shingle: Dental ethics and search engine optimization." Elsevier, 2018.

[33] "Attracting and analyzing spam postings," Nov. 2015.

[34] R. Hassanian-esfahani and M. Kargar, "Sectional MinHash for near-duplicate detection," Expert Syst. Appl., vol. 99, pp. 203–212, Jun. 2018, doi: 10.1016/J.ESWA.2018.01.014.

[35] R. Agrawal, "Controlling Unethical Practices in Web Designing by Search Engines," 2018.

[36] J. J. Whang, Y. S. Jeong, I. S. Dhillon, S. Kang, and J. Lee, "Fast Asynchronous Anti-TrustRank for Web Spam Detection," 2018.

[37] D. Pawade, A. Sakhapara, M. Jain, N. Jain, and K. Gada, "Story Scrambler–Automatic Text Generation Using Word Level RNN-LSTM," Int. J. Inf. Technol. Comput. Sci., vol. 10, no. 6, pp. 44–53, 2018.

[38] W. Li, "Consistency checking of natural language temporal requirements using answer-set programming," 2015.

[39] K. McKeown, Text generation. Cambridge University Press, 1992.

[40] H. T. Dang, "Overview of DUC 2005," in Proceedings of the document understanding conference, 2005, vol. 2005, pp. 1–12.

[41] A. Summerville, S. Snodgrass, M. Guzdial, C. Holmgard, A. K. Hoover, A. Isaksen, A. Nealen, et al., "Procedural Content Generation via Machine Learning (PCGML)," IEEE Trans. Games, vol. 10, no. 3, pp. 257–270, Sep. 2018.

[42] D. Roy, M. Mitra, and D. Ganguly, "To Clean or Not to Clean," J. Data Inf. Qual., vol. 10, no. 4, pp. 1–25, Oct. 2018.

[43] E. Sadredini, D. Guo, C. Bo, R. Rahimi, K. Skadron, and H. Wang, "A Scalable Solution for Rule-Based Part-of-Speech Tagging on Novel Hardware Accelerators," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18, 2018, pp. 665–674.

[44] A. Keyaki and J. Miyazaki, "Part-of-speech tagging for web search queries using a large-scale web corpus," in Proceedings of the Symposium on Applied Computing - SAC '17, 2017, pp. 931–937.

[45] Y. S. Toutanova, Kristina, Dan Klein, Christopher Manning, "Feature-rich part-of-speech tagging with a cyclic dependency network," in In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for computational Linguistics, 2003, pp. 173–180.

[46] X. Dai, N., Davison, B. D., & Qi, "Looking into the past to better classify web spam," in In Proceedings of the 5th international workshop on adversarial information retrieval on the web, 2009, pp. 1–8.

Combining 3D Interpolation, Regression, and Body Features to build 3D Human Data for Garment: An Application to Building 3D Vietnamese Female Data Model

Tran Thi Minh Kieu[1]
School of Textile – Leather and Fashion, Hanoi University of Science and Technology, Hanoi, Vietnam

Nguyen Tung Mau[2]
School of Textile – Leather and Fashion
Hanoi University of Science and Technology
Industrial University of Hochiminh City
Hochiminh City, Vietnam

Le Van[3]
Information Science Faculty
Sai Gon University, Hochiminh City, Vietnam

Pham The Bao[4]
Information Science Faculty
Sai Gon University
Hochiminh City
Vietnam

***Abstract*—Modeling 3D human body is an advanced technique used in human motion analysis and garment industry. In this paper, we propose a method for forming deformation functions so that we can rebuild the 3D human body given anthropometric measurements. The advanced idea in our approach is that we split the 3D body into small parts. In that way, we can specialize different set of parameters needed to interpolate for each section. With an interpolation approach, we build a 3D human body for 593 female bodies with the corresponding body shape but require fewer input measurements than 3D laser scans.**

Keywords—Anthropometry; 3D scanning; human body modeling; interpolation; parametric modeling

I. Introduction

With the rapid increase in 3D technology, 3D human models have more practical applications than ever. In the field of animation, 3D virtual characters used in films or videos are much similar to a real human. Hologram call will be the next step for our telecommunications industry and especially for modern life. When people need more time for their family and work, virtual try-on technology or custom suit online will be cut off the time for shopping. However, full body 3D scanners nowadays are still unaffordable for individual [4]. Besides, scanning systems usually are complicated to set up, and there is rarely a friendly manual.

Those reasons have fueled us to propose a more convenient method for non-expert users, who want a 3D body ourselves without the 3D scanner. With the slightly greater number of key points compared to those used in traditional manners, but much smaller than those used in 3D body scanner, we can estimate 3D human models simply by using a personal computer. To achieve this work, we choose n based-slices to present the whole human body, then fine-tune slice-by-slice to adapt to any particular human body shapes. Finally, meshes are used to connect the entire body.

We rely on tailor's (circumference) measurements at specific places on the human body to construct essential shapes. We build data points with their respective shapes and perimeters to get important slices. From these data slices, we use interpolation to estimate intermediate slices that form the entire 3D body of a human. The dataset used in our experiment contains 593 3D women models which are saved under "obj" format; each model has a set of measurements that is the input of our method. This dataset is from the garment research group of Hanoi University of Technology. We use relative error to compute the error level of the proposed approach; the average square error is 11.02621202.

This research was supported by the Hanoi University of Science and Technology (HUST) under project number T2018-PC-048.

II. Related Works

Many techniques have been developing in the field of 3D modeling. The following approaches are vital to the 3D garment design [2].

- Parametric learning: this method creates a human template and then combines with semantic features from an individual to produce a new human model [5].

Wang et al. [1] proposed a method to create a parametric model based on the vital body sizes and Spline interpolation. To satisfy the garment or animation's requirement, the author presented two modules: the skeleton fitting and independent cross sections for each body part. After having a surrogate model, the users can manipulate the key dimensions of the surrogate model to generate their desired 3D model. Cordier et al. [3] developed an Online Clothing Store as a web application. The application requested eight primary measurements of the human body as an input to a generic model. Deforming existing model in a database will be easier to obtain a new one.

- Statistics-Based Synthesis Model: This method generates a human body based on the investigation of the body shape distribution. However, this method does not have a true technique to build a new individual model from the input constraints [1, 9].

III. Methodology

A. Overview

In this paper, we use the Hermite Curve Interpolation method to interpolate our model. Hermite curves are widely used in computer graphics and geometric models to obtain motion curves or orbits that pass through certain points of the plane or three-dimensional space. In these applications, each coordinate of the plane or space is interpolated using parameters which represent the tension of the curve. We use this method because the curve properties on the human body [9].

In Hermite's method, the third is determined by the first and last points with two tilt angles at those two points. Hermite curves are often used to interpolate data at given values $x_1, x_2, \ldots, x_n$ to achieve a "smoth" continuous function. The data includes the value of the desired function and the derivative at each x_k. The Hermite formula is applied to each interval (x_k, x_{k+1}) separately. Polynomial interpolation of the Hermite curve is given in formula (1), and the example is shown in figure 1.

$$p(t) = m_1 h_3(t) + p_1 h_1(t) + p_2 h_2(t) + m_2 h_4(t) \qquad (1)$$

with m_1, m_2, p_1, p_2 are tangential values and $h_1(t), h_2(t), h_3(t), h_4(t)$ are the basic Hermit functions corresponding to

$$h_1(t) = 2t^3 - 3t^2 + 1$$

$$h_2(t) = -2t^3 + 3t^2$$

$$h_3(t) = t^3 - 2t^2 + t$$

$$h_3(t) = t^3 - 2t^2 + t$$

The Hermite curve with tension parameters is formulated as (2).

$$p(t) = \begin{bmatrix} t^3 & t^2 & t & 1 \end{bmatrix} \begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ m_1 \\ m_2 \end{bmatrix} \qquad (2)$$

with $m_1 = s(p_2 - p_0)$ and $m_2 = s(p_3 - p_1)$

Then we have (3),

$$p(t) = \begin{bmatrix} t^3 & t^2 & t & 1 \end{bmatrix}$$

$$\begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -s & 0 & s & 0 \\ 0 & -s & 0 & s \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ p_2 \\ p_3 \end{bmatrix} \qquad (3)$$

We generalize to (4).

$$p(t) = \begin{bmatrix} t^3 & t^2 & t & 1 \end{bmatrix} \begin{bmatrix} -s & 2-s & s-2 & s \\ 2s & s-3 & 3-2s & -s \\ -s & 0 & s & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_{k-1} \\ p_k \\ p_{k+1} \\ p_{k+2} \end{bmatrix} \qquad (4)$$

B. Building Model

1) General model: Diagram 1 shows our proposed framework to generate a 3D human model given a set of hand measurements.

2) Building a set of measurements for the model: The Standard measurements needed to make clothes are listed in Table 1 and 2; these are fundamental measures used in the field of costume design [3, 6].

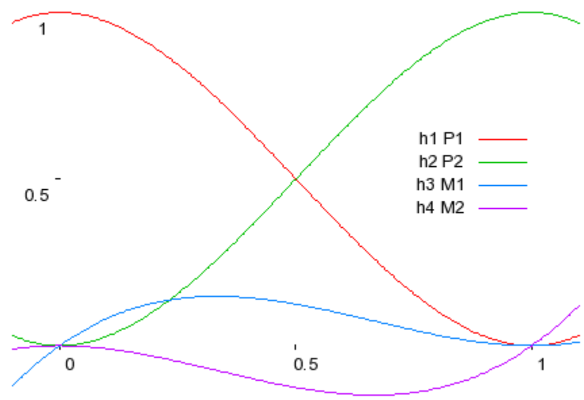


Fig. 1. The basic Hermit Functions.

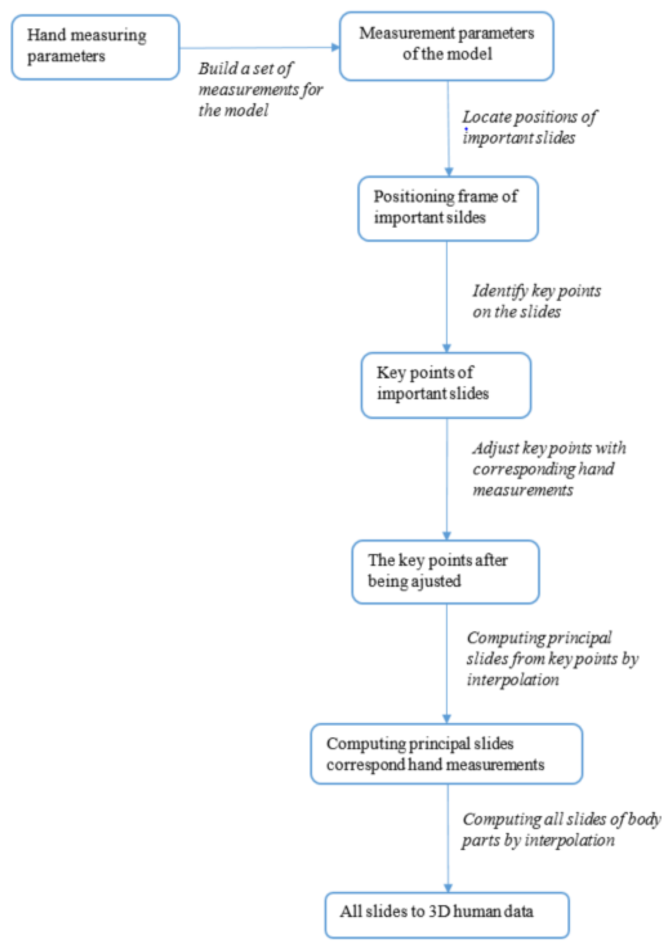


Diagram. 1. General Model.

TABLE. I. BASIC MEASUREMENTS FOR SEWING CLOTHES

No	Measure names	Measurement methods
1	Shoulder	Measure the distance from the tip of the shoulder to the side of the shoulder and the other shoulder
2	Shirt length	Measure the distance from the 7th vertebrae of the spine to the level of the perineum
3	Short sleeve length	Measure the distance from the shoulder to the elbow
4	Long sleeve length	Measure the distance from the shoulder to the wrist
5	Necklace	Use a tape to measure the circumference around the neck
6	Chest circumference	Use a tape to measure the circumference of the chest
7	Waist circumference	Use a measuring tape to measure the circumference of the belly circumference above the navel 2 cm

TABLE. II. BASIC MEASUREMENT FOR SEWING PANT

No	Measure names	Measurement methods
1	Long bottom	Measure the distance from the waistline to the bottom of the perineum
2	Short pants	Measure the vertical distance from the perineum to the level of the knee
3	Long pants	Measure the distance from the perineum to the ground
4	Waist circumference	Use a measuring tape to measure the circumference of the belly circumference above the navel 2 cm
5	Buttocks	Use the tape to measure the circumference of the butt ring at the most protruding level of the large transfer tab
6	Middle thigh	Use a measuring tape on the circumference of the thigh, not tighten the ruler, measure the middle half between the hip and the knee
7	Calf ring	Use the horizontal measuring tape to circle the calf of the most enlarged calf
8	Leg ring	Use a tape measure to measure the horizontal circumference of the anklet

From the above basic measurements, we build a set of measuring parameters to create a suitable 3D model in the field of costume design. This set of parameters includes 9 measurements of the circumference of the body, and 7 measurements of body length/width are shown in Table 3 [1, 7].

3) Building a model to locate the position of pricipal slices: The principal slices are at positions that represent the body measurements of the 3D model; we locate those slices to interact with the hand measurements. From measurements of the length of the set of parameters, we can locate the primary positions of the body parts such as calf, thigh, buttocks, abdomen, chest, and neck. It is also easy to locate the slices to be used to change the measurement in each region. A set of frames being interacted with is constructed as shown in Figure 2, and the meanings are explained in Table 4. We build algorithm 1 to locate the position of principal slices.

TABLE. III. MEASUREMENT PARAMETERS FOR 3D MODELS ARE BUILT

No	Measure names	Measurement methods
1	Necklace	circumference
2	Chest ring	
3	Waist circumference	
4	Buttocks	
5	Middle thigh	
6	Calf circumference	
7	Leg circumference	
8	Round the upper arm	
9	Wrist cuff	
10	Length from nape to belly	length/width
11	Length from belly to perineum	
12	Thigh length	
13	Calf length	
14	Upper arm length	
15	Lower arm length	
16	Shoulder width	

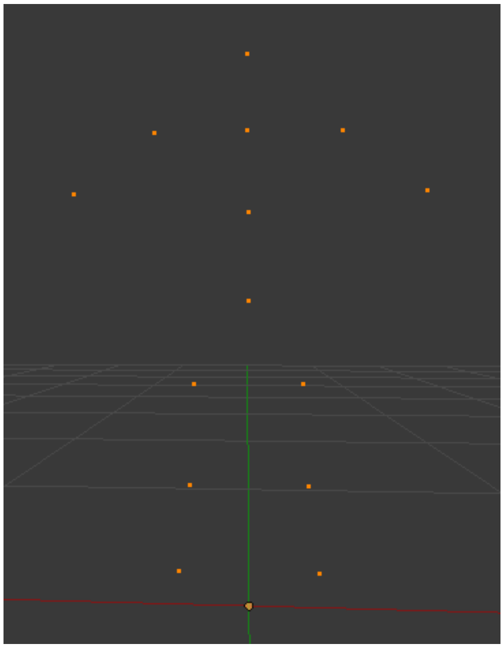


Fig. 2. The Positioning Frame of Principal Slices will be Interacted.

TABLE. IV. MEANING OF POSITIONS

Position	Meaning
(1)	Location of the cross section
(2)	Chest circumference position
(3)	The position of the abdomen cross section
(4)	The position of the buttocks section
(5)	The position of the middle thigh section
(6)	Face position cuts the calf ring
(7)	The position of the cut leg face
(8)	The cross-sectional position of the upper arm
(9)	The cross-sectional location of the wrist

Algorithm 1 Locating the position of principal slices

```
Algorithm to locate the position of principal slices.
Input: the hight of human – H, and data of model – M.
Output: the set of positions of principal slices (It were arranged), P.
P = ∅;
Set center = (0, 0);
for (i=1 to M.getNumber()) do
    P[i].setPosition(H*M[i].getRatioOfPosition());
end for
return P;
End
```

4) Building a model that identifies the key points of each slice: Each human body can be modeled simply by a set of slices, and each slice will contain points to form a curve. To create a controllable human body model, we need to identify some key points on each slice and use these points to create a curve. To find the key points of each slice, we improved the technique which was introduced by Shuxia Wang [1] as follows:

- Kind 1: all slices on the body parts (except the body) have a nearly circular structure, so we will choose 4 extreme points to make the key points of the section, Figure 3 and 4.
- Kind 2: the slices on the body have the complex structure; we observe the symmetry and make statistics to find the 12 key points representing the slice as shown, Figures 5 and 6.

we designed algorithm 2 to solve this problem.

Algorithm 2 Determine the kind for each principal slice.

```
Algorithm to determining the kind for each principal slice
Input: the hight of human - H, the ratio of principal slices with hight (using
statistic to computing from data) – Ra, and positions of slices – P.
Output: the kind and center positions of all principal sildes, PS
Set center = (0, 0);
Set PS = ∅;
for (i=1 to P.getNumber()) do
    PS[i].setPosition(P[i].getPosition());
    PS[i].setKind(Ra[i].getKind());
    PS[i].setPositionX(H*Ra[i].getRatioX());
    PS[i].setPositionY(H*Ra[i].getRatioY());
    PS[i].setPositionZ(H*Ra[i].getRatioZ());
end for
return PS;
End
```

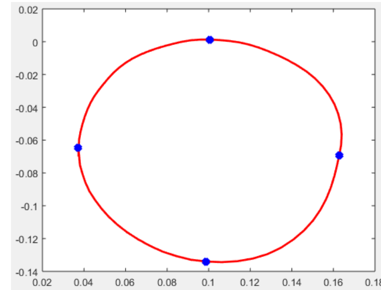


Fig. 3. The Key Points of the Thing Slice.

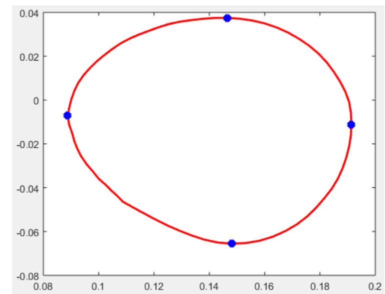


Fig. 4. The Key Points of the Calf Slice.

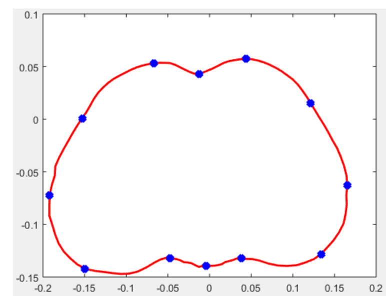


Fig. 5. The Key Points of the Buttocks Slice.

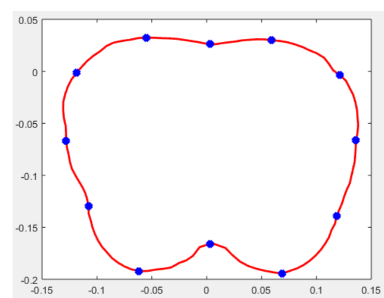


Fig. 6. The Key Points of the Chest Slice.

5) Building a model to adjust the key points on the slice corresponding to hand measurements: Once having the key points which represent each slice, we adjust those points to fit the slice shape with the use of approximate circumferences based on hand measurements, algorithm 3. We conduct the following steps:

- Step 1: Determine the center, and compute the ratio of the large radius, and small radius of the current slice of the model, Figure 7.
- Step 2: Find the large radius of the current slice of the model (Radius = Perimeter / (2 * π)), Figure 8.

- Step 3: Compute the large radius of the real slice based on the hand measurement parameters (Radius = Measure / (2 * π)), Figure 9.
- Step 4: Adjust the points corresponding to the large radius of the new slice, Figure 10.
- Step 5: Adjust the point on the small radius in proportion to the large radius of the current slice, Figure 11.

Algorithm 3 Adjusting the key points of one principal slice.

Algorithm to adjust the key points
Input: perimeter (pe) of slice (hand measurement), the set of key points, P
Output: the set of key points after it were adjusted, P
Set center = (0, 0);
$\text{Radius}_{\text{small}}$ = P.getSmallRadius();
$\text{Radius}_{\text{large}}$ = P.getLargeRadius();
Ratio = $\text{Radius}_{\text{small}}/\text{Radius}_{\text{large}}$;
$\text{Radius}_{\text{measure}}$ = pe/(2*π); // R is the large radius of the circle
$\text{Ratio}_{\text{model and real}}$ = $\text{Radius}_{\text{measure}}$/ $\text{Radius}_{\text{large}}$;
for (i=1 to P.getNumberOfSet()) do
    if(P[i] is on large radius) then
        P[i].adjustPoint($\text{Ratio}_{\text{model and real}}$);
    end if
    if(P[i] is on small radius) then
        P[i].adjustPoint($\text{Ratio}_{\text{model and real}}$, Ratio);
    end if
end for
return P;
End

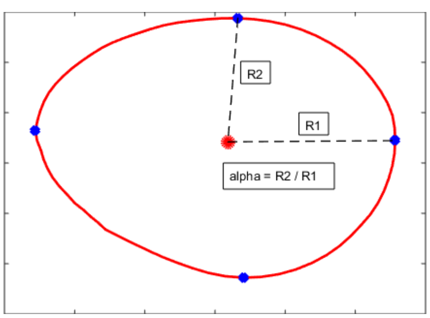


Fig. 7. Step 1 of the Solution.

6) Building the principal slice from key points by interpolation: After having key points on each principal slice, we interpolate more points to create the curve corresponding to the slices. We use Hermite Curve Interpolation method to perform interpolation because of three following reasons:

- This curve can pass through key points of the slice smoothly and without corners.
- This curve does not need to include tangential values at points. Its slope at each point will be computed from the coordinates of the two adjacent points.
- This curve can be controlled by a tension parameter, with different tension parameter values creating different curves.

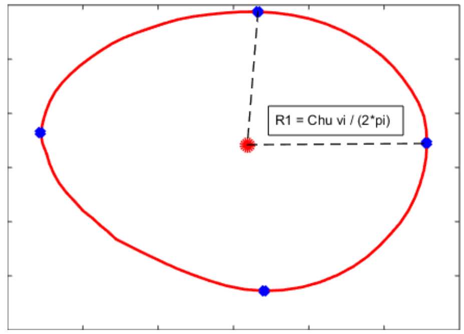


Fig. 8. Step 2 of the Solution.

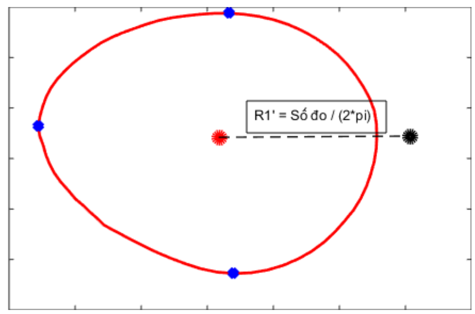


Fig. 9. Step 3 of the Solution.

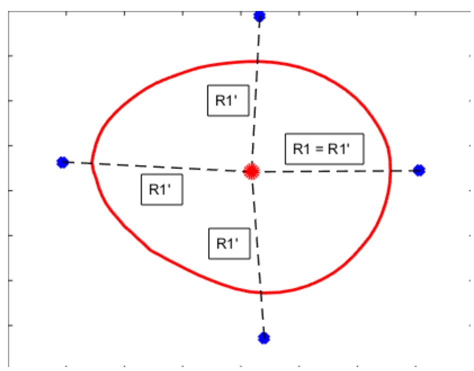


Fig. 10. Step 4 of the Solution.

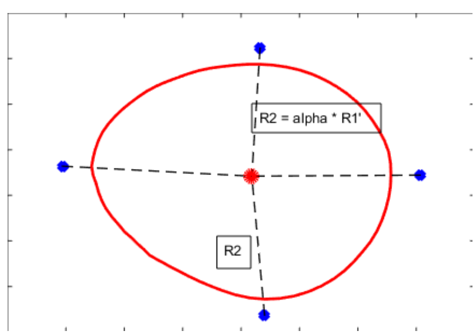


Fig. 11. Step 5 of the Solution.

Based on the experimental results as well as the results of Shuxia Wang [1], the tension parameter of -0.4 is suitable for the curve of the human body profile. The appropriateness is shown in the figures depicted below with the red "o" points as the raw data points on the slice scanned from the 3D Scanner, the red curve with the tension parameter of 0.8, the blue curve has a tension parameter of -1.6 and a black curve with a tension parameter of -0.4, Figure 12, 13. We use Algorithm 4 to generate data points for the principal slices for the body, Figure 14.

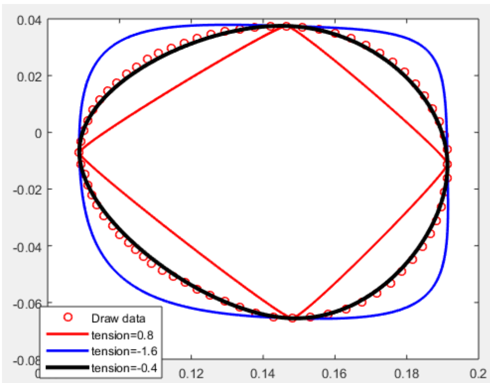


Fig. 12. Interpolation Curvers from 4 Key Points with different Tension Parameters.

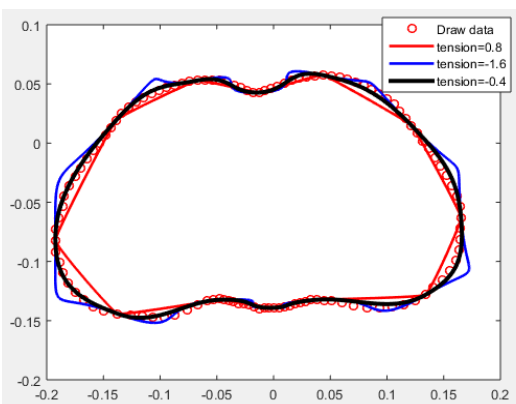


Fig. 13. Interpolation Curves from 12 Key Points with different Tension Parameters.

Algorithm 4 Generating all points (data) from key points by Hermite Curve Interpolation method.

```
Algorithm generation points
Input:     the set of key points – P, tension parameter – s.
Output:    all points were generated – D.
N = P.getNumber();
for (i=1 to N) do
           temp = generatePointsByHermitMethod(P[i], P[(i+1) mod N], P[(i+2) mod N], s);
           D = D ∪ temp;
end for
return D;
End
```

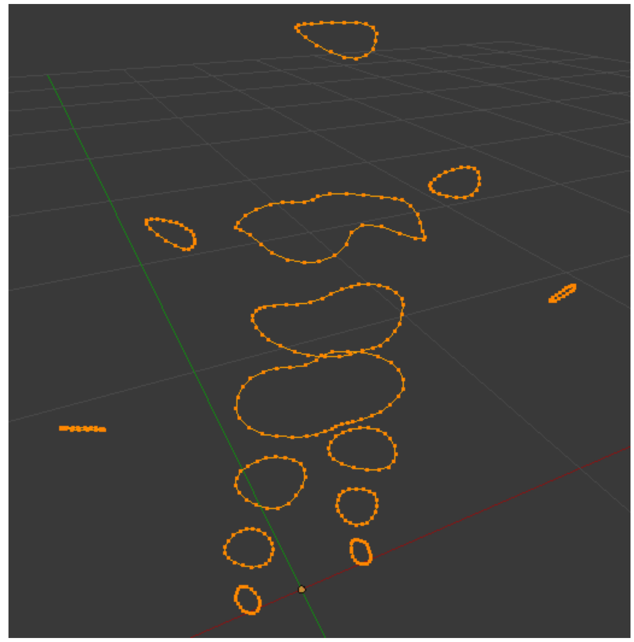


Fig. 14. The Principal Slices on the Body after Interpolation.

7) Building interpolation models from principal slices on parts of the body: Based on the constructed principal slices, we generate extra slices in different parts of the body. The greater the number of slices of each body part, the more detailed it will be, with the tradeoff computational costs. In this research, we chose the number of sections that are sufficient to represent for parts of the body, as shown in Table 5. We use Algorithm 5 to generate these intermediate slices. And result for Algorithm 5 as Figure 15, 16, 17, 18, 19.

TABLE. V. NUMBER OF SLICES FOR EACH REGION

Region	Number of slices	Kind
Chest	6	1
Stomach	10	1
Buttocks	6	1
Neck	9	1
Thighs	14	2
Calf	10	2
Leg	3	2
Upper arm	12	2
Lower arm	8	2
Wrist	5	2

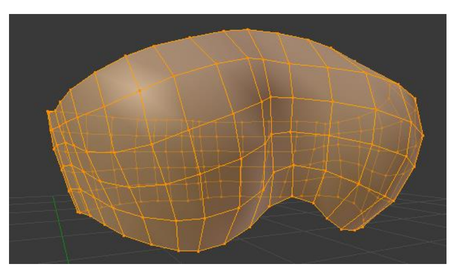


Fig. 15. The slices of the chest region

Algorithm 5 Generating internal slices of regions.

```
Algorithm to generate internal slices of regions
Input:      principal slices - PS
Output:     data of 3D human model – Re.
Re = ∅;
for (i=1 to PS.getNumber()) do
   N = PS[i].getNumber();
   if(PS[i].getKind()==1) then
      for (j=1 to N) do
         temp = generateInternalSlices(PS[i].getPrincipalSlice(),
PS[i].getPosition() + j);
         Re = Re ∪ temp;
      end for
   end if
   if(PS[i].getKind()==2) then
      for (j=1 to N) do          // generate data for left side
         temp =
generateInternalSlices(PS[i].getPrincipalSlice(),PS[i].getPosition() + j);
         Re = Re ∪ temp;
      end for
      for (j=1 to N) do          // generate data for right side
         temp =
generateInternalSlices(PS[i].getPrincipalSlice(),PS[i].getPosition() + j);
         Re = Re ∪ temp;
      end for
   end if
end for
End
```

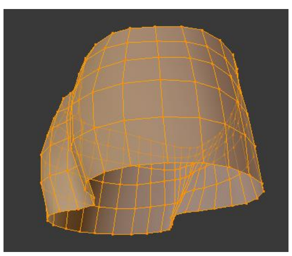


Fig. 16. The Slices of the Buttocks Region.

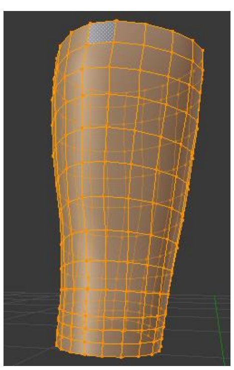


Fig. 17. The Slices of the thing Region.

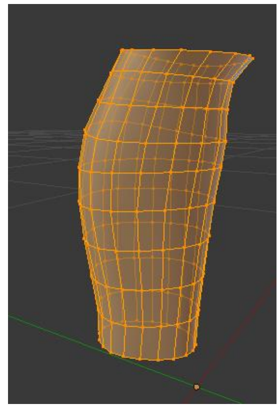


Fig. 18. The Slices of the Calf Region.

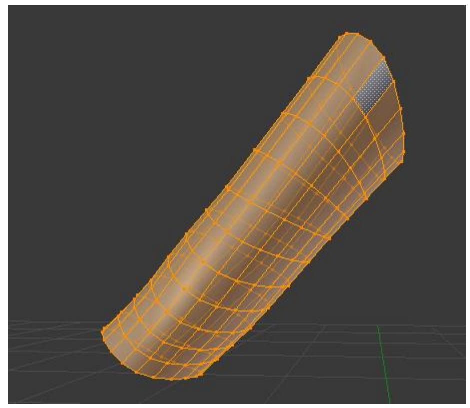


Fig. 19. The Slices of the upper Arm Region.

C. Error Evaluating Method

To evaluate errors, we compute the circumference of the slices from the 3D built model. Formula (5) shows how we compute the circumference of slice ith.

$$p_i = \sum_{j=1}^{n-1} \sqrt{\left(x_{i,j} - x_{i,j+1}\right)^2 + \left(y_{i,j} - y_{i,j+1}\right)^2 + \left(z_{i,j} - z_{i,j+1}\right)^2} + \sqrt{\left(x_{i,n} - x_{i,1}\right)^2 + \left(y_{i,n} - y_{i,1}\right)^2 + \left(z_{i,n} - y_{i,1}\right)^2} \quad (5)$$

where n is the number of points in slice ith and $(x_{i,j}, y_{i,j}, z_{i,j})$ are the coordinates of the jth poi nt at slice ith.

$p_{groundtruth}$ is the circumference of the real data scanned by the 3D scanner, $p_{prediction}$ is the circumference of slice of the 3D model that we construct from the hand measurements, and n is the number of models. There are two types of errors that we consider, they are the errors of principal slices and the errors of 3D model.

1) Evaluating the errors of principal slices: The errors of principal slices are the errors that we only calculate at principal slices representing the body's measurements, Formula 6

$$SizeLoss = \frac{\sum_{i=1}^{n}\left(p_{\text{groundtruth, i}} - p_{\text{prediction,i}}\right)^2}{n} \quad (6)$$

2) Evaluating the errors of 3D built model: The error of the 3D built model is the error which is computed by slices the whole body of the 3D built model, formula (7).

$$BodyLoss = \frac{\sum_{i=1}^{n} BodyLossPerOneModel}{n} \qquad (7)$$

with $BodyLossPerOneModel = \frac{\sum_{i=1}^{m} (p_{groundtruth,i} - p_{prediction,i})^2}{m}$

where m is the number of slices of the whole body.

IV. Experiments and Results

A. Data and Experimental Environment

We conduct experiments on the data set of 593 3D scans of women body; each model has a set of manual measurements, and its point cloud is saved under ".obj" format. This dataset was recorded by the garment research group of Hanoi University of Technology [8].

We use Python as the main programming language. The algorithms are implemented on a computer using CPU Intel Core i5, and 4GB RAM. After generating the 3D model, we utilize Blender for visualizing.

B. Results

1) Program simulating 3D mannequins corresponding to hand measurements: We created a program that simulates the female 3D mannequin corresponding to the manual parameters given by users, as shown in Figure 20. The program helps to observe the actual 3D model and the generated model, Figure 21.

2) Errors: The results of the square error of the principal slices and the squared error of the 3D built model are shown in Table 6 and Table 7.

There were 95 models in the dataset, which were damaged when scanning. These damaged models make the max square error significantly increase, and that leads to the high average error. Figure 22 shows a female 3D model from a defective 3D scanner. Table 8 and Table 9 show the results after eliminating faulty 3D data.

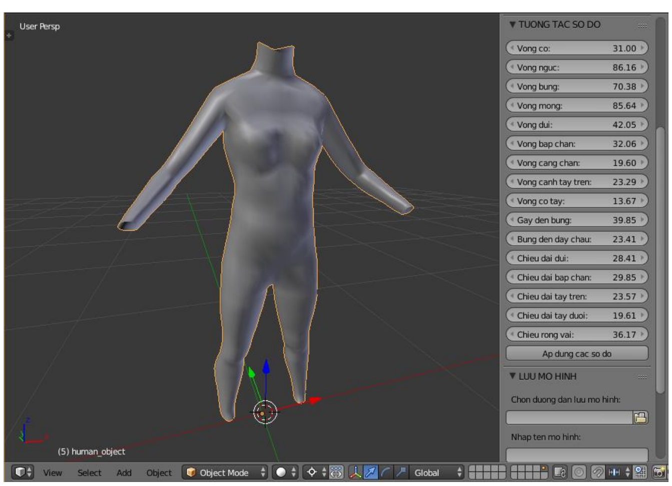


Fig. 20. Simulating 3D Mannequins Corresponding to Hand Measurements.

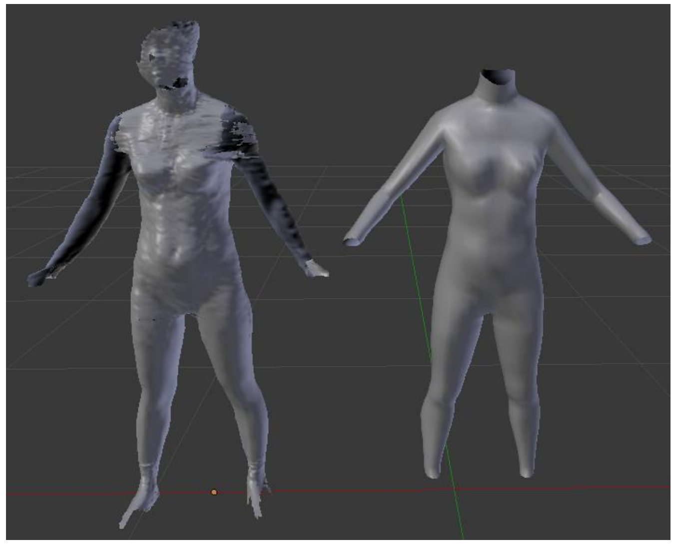


Fig. 21. Female 3D Model: (Left) from 3D Scanner; (Right) Built by Proposed Method.

TABLE. VI. Errors of Principal Slices

Measuring name	**Mean square error**
Necklace	2.6129091253
Chest ring	9.1809139148
Waist circumference	0.9392992822
Buttocks	1.0371366978
Middle thigh	0.0371047278
Calf circumference	0.0000332125
Leg circumference	0.0000247387
Upper arm	1.0321699121
Wrist cuff	2.0574341568

TABLE. VII. Errors of 3D Built Model

Mean square error	11.02621202
Min square error	1.38504
Max square error	125.55837

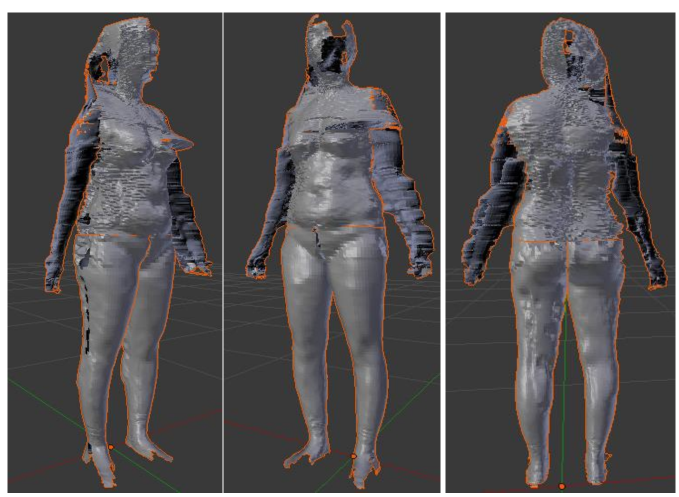


Fig. 22. The Female 3D Model from the 3D Scanner is Wrong.

TABLE. VIII. ERRORS OF PRINCIPAL SLICES AFTER SEPARATION OF FAULTY DATA

Measuring name	Mean square error	
	498 good data	**95 faulty data**
Necklace	0.00015151	16.3092595715
Chest ring	0.0001203749	57.3076000503
Waist circumference	0.0000326014	5.8630340931
Buttocks	0.0000614009	6.4735945702
Middle thigh	0.0000339809	0.2314334855
Calf circumference	0.0000394556	0.00004396
Leg circumference	0.0000294007	0.0000372449
Upper arm	0.0000224071	6.4427957809
Wrist cuff	0.0000610313	12.8424006462

TABLE. IX. ERRORS OF 3D BUILT MODEL AFTER SEPARATION OF FAULTY DATA

	498 good data	**95 faulty data**
Mean square error	5.491083893	23.52586951
Min square error	1.38504	10.06982
Max square error	9.95929	125.55837

V. CONCLUSION

In this study, a method of building 3D human models is presented. We introduce a framework to decide which primary slices needed to interpolate before generating the entire 3D model. We conduct the experiments on the dataset of 3D scans of female body and investigate a set of appropriate measurements that used as the input. The proposed method mainly based on the Hermite interpolation, which has a closed form and does not require many calculations, results in a small computational cost.

The main drawback of this approach is that highly depends on the selection of primary slices. Interpolating the intermediate slices only depending on the primary slices of a particular model made a generated model seems reasonable but not practical because of the complexity of the body structure. In practice, However, we rarely have manual measurements at all positions on the human body. One novel way to overcome this drawback is using Deep Learning approach, which takes advantage of information containing in the dataset to infer the intermediate slices.

REFERENCES

[1] S. Wang, S. Qin and C. Guan, "Feature-Based Human Model for Digital Apparel Design," IEEE Transactions on Automation Science and Engineering, vol. 11, no. 2, pp. 620 – 626, 2014.

[2] Y. J. Liu, D. L. Zhang, and M. M. F. Yuen, "A survey on CAD methods in 3D garment design," Comput. Ind., vol. 61, pp. 576-593, 2010.

[3] F. Cordier, S. Hyewon, and N. Magnenat-Thalmann, "Made-to-measure technologies for an online clothing store," IEEE Comput. Graph. Appl., vol. 23, pp. 38-48, 2003.

[4] S. Y. Baek and K. Lee, "Parametric human body shape modeling framework for human-centered product design," *Comput.-Aided Design*, vol. 44, pp. 56–67, 2012.

[5] H. Seo and N. Magnenat-Thalmann, "An example-based approach to human body manipulation," *Graph. Models*, vol. 66, pp. 1–23, 2004.

[6] Jihong Xu, Wenbin Zhang, "The vacant distance ease relation between body and garment," Second International Conference on Information and Computing Science, Manchester, pp. 38-41, 2009.

[7] Park Soonjee, Miyoshi Machiko, "Development of Theory and Auto CAD Program for Designing the Individual Bodice Pattern from 3D Scanning Data of Human Body", Journal of ARAHE, 10.4, pp.216-225, 2003.

[8] Tran Thi Minh Kieu, Soonjee Park, "Development 'Aodai' pattern for Vietnamese women using 3D scan data", Korean Society for Clothing Industry (KSCI) International Conference, pp.390-391, 2011.

[9] Kochanek D. and Bartels R., "Interpolating Splines with local tension, continuity, and Bias control," Proceedings of the 11th International Conference on Computer Graphics and Interactive Techniques, ACM Press, pp. 33 – 41, 1998.

Breast Cancer Computer-Aided Detection System based on Simple Statistical Features and SVM Classification

Yahia Osman[1], Umar Alqasemi[2]
Biomedical Engineering Program
Department of Electrical and Computer Engineering
King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia

***Abstract*—Computer-Aided Detection (CADe) systems are becoming very helpful and useful in supporting physicians for early detection of breast cancer. In this paper, a CADe system that is able to detect abnormal clusters in mammographic images will be implemented using different classifiers and features. The CADe system will utilize a Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) as classifiers. Adopting mammographic database from Mammographic Image Analysis Society (MIAS), for training and testing, the performance of the two types of classifiers are compared in terms of sensitivity, specificity, and accuracy. The obtained values for the previous parameters show the efficiency of the CADe system to be used as a secondary screening method in detecting abnormal clusters given the Region of Interest (ROI). The best classifier is found to be SVM showed 96% accuracy, 92% sensitivity and 100% specificity.**

Keywords—Breast cancer; MIAS; features extraction; SVM; mammogram; clusters; computer-aided detection systems; KNN; ROI

I. Introduction

Breast cancer is a disease occurred when the cells in the female breast grow randomly and out of control. The type of breast cancer depends on morphology and proliferation. A female human breast is made up of three main parts: lobules, ducts, and connective tissue. The lobules are the glands that produce milk. The ducts are the tubes that carry milk to the nipple. The connective tissue (which consists of fibrous and fatty tissue) surrounds and holds everything together [1]. Most breast cancers begin in the ducts or lobules.

Breast cancer can spread in later stages outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasized [1]. Breast cancer is known to be the most lethal among abnormal masses leading to deaths of 2.09 million women globally in 2018 according to World Health Organization (WHO) [2]. Recently, the survival rates have been increased due to more awareness about the disease from social media and more availability and advancement of healthcare technology especially mammography and other diagnostic imaging techniques [3]. Mammography is commonly used as a diagnostic imaging technique for detecting breast cancer due to its availability, less imaging duration, and lower cost than other methods such as Magnetic Resonance Imaging (MRI). On the other hand, Ultrasound Imaging is lower in cost but worse in terms of reproducible mapping to physical location.

In mammography, there are some factors that can lead to wrong decisions among the physicians, such as the appearance of microcalcifications. Furthermore, biopsy is painful for patients to support surgery decision. Hence, the use of CADe systems may ensure the decision without the need of biopsy. Our CADe system assumes known ROI by radiologist and supposed to aid at least as a secondary diagnosis method to support surgery decision.

II. Literature Review

Several previous studies have been published involving CADe system for breast cancer using mammography, contributed in presenting preprocessing algorithms, new features of more statistical significance or more relevant to the morphology of the abnormal images, and classifiers of better performance combined with set of features.

Arai et al. [4] separated the database taken from Japanese Society of Computer Aided Medical Imaging Technology into two parts, training and testing with the data proportion were 74% and 26%, respectively. The author used the features that are mostly statistical including mean, variance, max, coefficient of variation, standard deviation, and two additional features, 7 Hu moments and centroid. These features are extracted from Wavelet decomposition results of each detail, horizontal, approximation, diagonal, and vertical details. Support Vector Machine (SVM) classifier is used, and obtained sensitivity and specificity of 90% and 91.43%, respectively. This study included features obtained after image transformation that may complicate the training process of the classifier and it is more computationally expensive.

Khaoula et al. [5] proposed a Computer Aided Diagnosis system using Mini-MIAS database to detect the abnormal areas in digital mammograms, using only the dense breast category and classifies them into abnormal (benign and malignant) and normal. Then, electromagnetism-like (EML) optimization algorithm, followed by the edge-based detection algorithm FIS (Fuzzy Inference System) were used to identify the suspicious structures. As a result, the performance of this method with SVM classifier in terms of accuracy is 86.36%. The features used in this study are computationally expensive while accuracy attained is low.

Pratiwi et al. [6] found that Radial Basis Function Neural Network (RBFNN) is more accurate in classifying digital mammogram image with sensitivity of 97.22% and specificity of 91.49% for normal and abnormal classification (CADe), while in classifying benign and malignant lesions (Computer Aided Diagnosis or CADx), RBFNN's sensitivity is 100% and specificity is 89.47%. The author used features from Gray-level Co-occurrence Matrix (GLCM) and suggested that using another texture-based feature extraction, such as wavelet or curvelet, may be used in breast cancer classification in the purpose of improving the accuracy.

Setiawan et al. [7] studied the usage of Law's Texture Energy Measure (LAWS) features as descriptors for classifying mammogram images. Based on result of the experiment, LAWS features give better accuracy when classifying mammogram images compared to GLCM features. The true accuracy value of benign-malignant classification (CADx) is 78.21%, but using GLCM feature, the accuracy less than 55% for each degree. In this study, the author used ANN as classifier, suggested improvement can be done by changing the architecture of neural network model or by changing the number of nodes in the hidden layer.

Saad et al. [8] introduced an algorithm using Otsu's method for detection of Microcalcifications (MCs) and automatic diagnoses of breast cancer has been developed. The enhancement evaluation parameters such as contrast improvement index (CII), peak signal-to-noise ratio (PSNR), and Edge Preservation Index (EPI) conclude that enhancement algorithm significantly improved the contrast of MCs against the background and hence improved detection of MCs. The algorithm implemented also shows that adaptive boosting (Adaboost) classification is more sensitive and accurate for the detection of both single and clustered MCs as compared to the ANN [14]. The algorithm was tested for The Digital Database for Screening Mammography (DDSM), MIAS and local database and showed high level of overall accuracy (98.68%) and sensitivity (80.15%).

Pavel et al. [9] proposed a breast cancer detection method which uses Local Binary Patterns (LBP) features for breast representation. The proposed method was evaluated on a set created from MIAS and DDSM databases. The method showed accuracy close to 84% using SVM classifier only. This study used only LBP features which showed attractive accuracy [13]. The overall performance of the classifier can be improved if the ROI has been specified in this study.

Table. I summarize the previous studies involving breast cancer images using mammogram.

III. Database

MIAS is organized by U.K research groups that are interested in the understanding of mammograms and for image processing and recognition [10]. MIAS database consists of 322 images, which belong to three classes normal, benign and malignant. There are 208 normal, 63 benign and 51 malignant mammograms.

The detailed information about MIAS database included in an introduction file in seven information columns for each mammogram, for more information, refer to [11].

The dataset used in this study is part of MIAS database, includes 72 normal images non-cancerous and 72 abnormal ones cancerous (total 96 of which are used for training including 48 normal and 48 abnormal, and total 48 of which are used for testing including 24 normal and 24 abnormal) (diagnostic details of abnormal cases are shown in Table II). The software used in this study is MATLAB V2019b network licensed through the university system.

TABLE. I. Previous Studies Summarization

Author and date	**Features Used**	**Features Elimination Technique**	**Classifiers**
Arai et al. [4]	Max, Mean, Variance, STD, CV, Centroid, 7 Hu, and Wavelet	N/A	SVM
Khaoula et al. [5]	FIS and Zernike Moments	N/A	SVM
Pratiwi et al. [6]	GLCM (ASM, Correlation, Sum Entropy, and Sum Variance)	T-test	Back-PNN and RBFNN
Setiawan et al. [7]	Laws' texture, energy measures, and GLCM	T-test for GLCM only	ANN
Saad et al. [8]	LAWS, GLCM, Kurtosis, and Skewness	N/A	ANN and Adaptive boosting
Pavel et al. [9]	LBP	N/A	SVM
Author and date	**Accuracy**	**Sensitivity**	**Specificity**
Arai et al. [4]	84.44%	90.00%	91.43%
Khaoula et al. [5]	86.36%	81.81%	90.9%
Pratiwi et al. [6]	92.1%	97.22%	91.49%
Setiawan et al. [7]	93.90%	91%	100%
Saad et al. [8]	97.92%	64.33%	74.16%
Pavel et al. [9]	84%	------	------

TABLE. II. CLASS OF ABNORMALITY PRESENTS

Class of Abnormality	**Abbreviation**	**Number of Images**
Well-defined/circumscribed masses	CIRC	22
Speculated masses	SPIC	14
Architectural distortion	ARCH	11
Asymmetry	ASYM	13
Other, ill-defined masses	MISC	12

IV. METHODOLOGY

1) Preprocessing: Region of Interest (ROI) of 32x32 pixels is cropped around the marked center of the suspicious area marked by radiologist for all the dataset images. This is to reduce the computational load and to make feature computation more concentrated in the ROI not distracted by other details in the whole breast image [12]. During our study, we tried using the full size of the mammogram images (1024x1024), but the results were not significant.

2) Features Extraction: Initially we computed 94 features starting from the first order statistics (14 features) and texture features (64 Histogram features and 16 GLCM features). Then, using the T-test (significance p-value < 5%) and classifiers performance, the added features are eliminated manually after checking both the P-value of t-test and classifiers' performance parameters including accuracy, sensitivity and specificity. At the end, the final most contributing features used in this study after rounds of trial and error are first order statistical ones including mean, median, mode, and quantile (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9), those showed best t-test significance along with best classification performance.

3) Classifiers: The classifiers used in this study are shown in Table. III:

Fig. 1 illustrates the iterative steps used while designing the CADe system. Each block/step will be explained more within the following text.

TABLE. III. THE CLASSIFIERS USED

Classifier	**Abbreviation**	**Parameters**
Support Vector Machine	SVM	Linear, Polynomial, and Radial Basis Function
K-Nearest Neighbor	KNN	1, 2, 3, 4, and 5

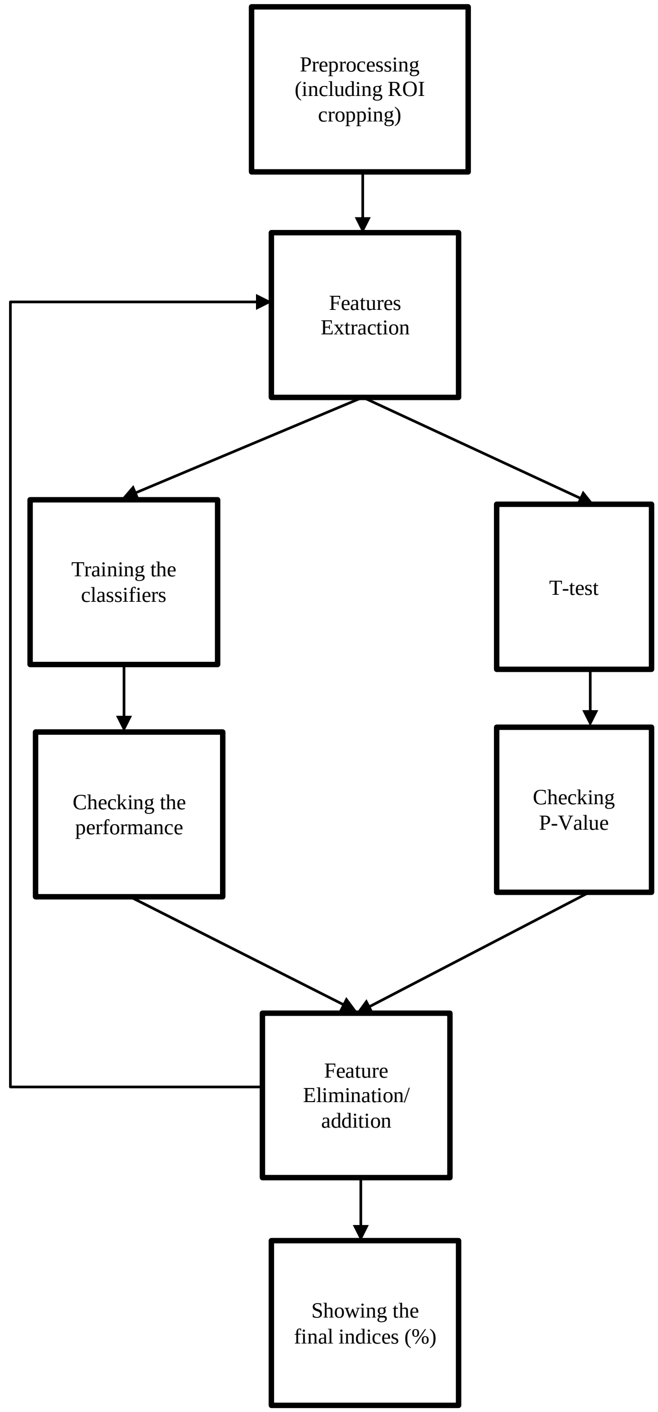


Fig. 1. Methodology Chart.

V. Results

The final results showed that the T-Test has a number of useful features (P-Value < 0.05) = 12 out of 12. Which means that all the used first order statistical features are significantly useful. The final results are shown in Table IV.

TABLE. IV. Final Results

Indices (%)	**SVM rbf**	**SVM Poly**	**SVM Linear**	**KNN 1**
Sensitivity	NAN	88%	92%	88.5%
Specificity	50%	91%	100%	95%
PPV	0%	92%	100%	96%
NPV	100%	87.5%	92%	87.5%
Accuracy	50%	89.5%	96%	92%
Error	50%	10%	4%	8%
Indices (%)	**KNN 2**	**KNN 3**	**KNN 4**	**KNN 5**
Sensitivity	86%	89%	86%	88.5%
Specificity	100%	100%	100%	95%
PPV	100%	100%	100%	96%
NPV	83%	87.5%	83%	87.5%
Accuracy	92%	94%	92%	92%
Error	8%	6%	8%	8%

Table IV shows that the best classifier was SVM-Linear with accuracy = 96% and Sensitivity = 92%. Followed by KNN-3 with an error that is equal to 6% only. The results in comparing to the previous studies were satisfying as the features used were only the simple first order statistics.

VI. Conclusion and Discussion

In this study, the final results were impressive in comparing to previous studies those used SVM classification. Given that, we used here simple first order statistics features. On the other hand, with Neural Network based classifiers, previous studies showed that computationally more expensive features gave comparable results to what we got here. Future studies can contribute by adding the microcalcifications (MCs) to the dataset (we excluded MCs in this study) and using the sophisticated classifiers, such as, ANN and RBFNN.

Table V shows the results of our study in comparing to the previous studies that used SVM classifier:

TABLE. V. Comparing Results with SVM Classifiers' Studies

Author and date	**Classifiers**	**Accuracy**	**Sensitivity**	**Specificity**
Arai et al. [4]	SVM	84.44%	90.00%	91.43%
Khaoula et al. [5]	SVM	86.36%	81.81%	90.9%
Pavel et al. [9]	SVM	84%	------	------
Our Study	SVM	96%	92%	100%

References

[1] D. o. C. P. a. Control, "Centers for Disease Control and Prevention," 11 September 2018. [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm. [Accessed 26 December 2019].

[2] W. H. Organization, "World Health Organization," 12 September 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed 26 December 2019].

[3] P. U. S. Alqasemi, Interviewee, Assistant Professor of Biomedical Engineering. [Interview]. 12 December 2019.

[4] I. N. A. H. O. R. K. Kohei Arai, "Improvement of Automated Detection Method for Clustered Microcalcification Based on Wavelet Transformation and Support Vector Machine," International Journal of Advanced Research in Artificial Intelligence, vol. 2, no. 4, pp. 23-28, 2013.

[5] M. N. S. A. T. Khaoula Belhaj Soulami, "A CAD system for the detection and classification of abnormalities in dense mammograms using Electromagnetism-like optimization algorithm," in 3rd International Conference on Advanced Technologies for Signal and Image Processing, Morocco , 2017.

[6] A. ,. J. H. ,. S. N. Mellisa Pratiwi, "Mammograms Classification using Gray-level Co-occurrence Matrix and Radial Basis Function Neural Network," in International Conference on Computer Science and Computational Intelligence, Indonesia, 2015.

[7] E.,.J. W. a. Y. P. Arden Sagiterry Setiawan, "Mammogram Classification using Law's Texture Energy Measure and Neural Networks," in International Conference on Computer Science and Computational Intelligence, Indonesia, 2015.

[8] A. K. Q. K. Ghada Saad, "ANN and Adaboost application for automatic detection of microcalcifications in breast cancer," The Egyptian Journal of Radiology and Nuclear Medicine, vol. 47, pp. 1803-1814, 2016.

[9] L. L. Pavel Kra´l, "LBP Features for Breast Cancer Detection," in International Conference on Image Processing, Arizona, 2016.

[10] M. A. S. Al-antari, "Computer-Aided Breast Cancer Detection and Diagnosis from Digital Mammograms," Faculty of Engineering, Cairo University Giza, Egypt, Egypt, 2015.

[11] A. F. Clark, "The mini-MIAS database of mammograms," 11 December 2012. [Online]. Available: http://peipa.essex.ac.uk/info/mias.html. [Accessed 25 July 2019].

[12] N. Petrick, H. Chan, B. Sahiner, and Datong Wei, "An adaptive density-weighted contrast enhance- ment filter for mammographic breast mass detection," Medi- cal Imaging, IEEE Transactions on, vol. 15, no. 1, pp. 59–67, 1996.

[13] A. Oliver, X. Llado´, J. Freixenet, and J. Mart, "False positive reduction in mammographic mass detection us- ing local binary patterns," MICCAI, pp. 286–293, 2007.

[14] A. AbuBaker Mass lesion detection using wavelet decomposition transform and support vector machine. International Journal of Computer Science & Information Technology. 2012 Apr 1;4(2):33.

EEG Emotion Signal of Artificial Neural Network by using Capsule Network

Usman Ali[1], Haifang Li[*,2], Rong Yao[3], Qianshan Wang[4], Waqar Hussain[5]
Syed Badar ud Duja[6], Muhammad Amjad[7], Bilal Ahmed[8]
College of Information and Compute, Taiyuan University of Technology, Taiyuan, Shanxi, China

***Abstract*—Human emotion recognition through electroencephalographic (EEG) signals is becoming attractive. Several evolutions used for our research mechanism technology to describe two different primaries: one used for combining the vital attribute, frequency sphere, and physical element of the EEG signals, and the architecture describes the two-dimensional image. Emotion realization is imposing effort in the computer brain interface field, which is mostly used to understand the field of education, medical military, and many others. The allocation issue arises in the required area of emotion recognition. In this paper, the allocation structure based on Caps Net neural network is described. The heder factor shows that the best point to classified the original EEG signals scarce group to using many of the algorithms like Lasso for a better function to used and other than occupy the heights.Furthermore, essential features like tiny subset take by input for the computer network attain for many ultimate emotional classifications. Many of the results show to alternate the best parameters model use and other network formats to making the Caps Net and another neural network act as the emotional valuation on EEG signals. It attains almost 80.22% and 85.41% average allocation efficiency under demeanor and view of the emotion pathway as compared to the Support Vector Machine (SVM) and convolutional neural network(CNN or ConvNet). A significant allocation edge attains the best conclusion and automatically enhances the performance of the EEG emotional classification. Deep learning access, such as CNN has widely used to improve primary allocation performance of motor symbolism-based brain-computer interfaces (BCI). As we know that CNN's limited allocation achievement degraded when an essential point data is distorted. Basically, in the electroencephalography (EEG) case, the signals consist of the same user are not measure. So we implement the Capsule networks (CapsNet), which is essential to extract many features. By that, it attains a much more powerful and positive performance than the old CNN approaches.**

Keywords—Emotion recognition; caps net; EEG signal; multidimensional feature; hybrid neural networks; CNN; Granger; motor imagery classification; deep learning

I. INTRODUCTION

Emotion is a sentimental state of human beings and plays a significant role in making decisions and social interactions. Most of the emotional pictures display out the human nerves system and other mental systems to sense and understand human emotion with the help of the computer system, especially in human-computer interaction. Many machines are able to present the human emotion and other nerves terminology. In a present research method, emotion realization is mostly separated by the other two classes[26]. One thing is the critical castigatory depends on the non-physiological signal such that the speech recognition body of the human attitude and other facial expressions. However, other necessary and essential thing is to use physiological signals like myoelectric signal ECG, EEG, and its signal are mostly attaining the cerebral codex's changes directly reflect the human emotion in the current state. The basic research approach of emotion reorganization for the important branches divided into two different classes[14]. Class I is used to show the fundamental category depends on the signal and other non-physiological states like speech, body attitude, and the other face term. Another necessary and essential thing is used physiological signals like the myoelectric, ECG, EEG beside these EEG signals are directed to attain the deep layer because all the other human emotions will be change by reflecting straight[8].

Fig. 1 shows the architecture of the CapsNet. EEG used for divided into more emotional signals. In this paper, one Caps Net neural network used for priority. Stained flow scheme mostly used for making Caps Net neural network structure to imagine the EEG emotion sensor and comparison to classified by using another relevant algorithm. Now a days, different approaches are used such as the sense-based control system and track approaches to increase the collaboration with other devices. Brain-computer interface (BCI) is the best technology used for defect people[24]. So, this technology is used to provide a comfortable environment to improve the defect person. Brain-computer interface (BCI) tented approaches to deliver the technology to help defect person indirectly by using the mechanism of the machine without any other physical cooperation to transform by the electroencephalogram achievement. The nerves system of the human called neurotransmission movement used for supervising the direct way of the sequential preprocessing. In this paper, we used the eradication and transcription procedure, which is generally used in Brain-computer interface (BCI) in EEG signal and SMR appears whenever motor symbolism and movement[1]. A fundamental method is discuses to apply any Cap-Net to divide two major classes used for the EEG signal for imagery motor [11]. We always have seen that Caps Net escaped about the conflicting signal that used for the symbolism motor, and the other is determined. If we want to make the efficient architecture of the Caps Net to sort the signals. This technique is precious using the primary signals to see the clear image of the EEG that is converted 2D pictures to other STFT algorithms using the necessary convert signals to the domain time-frequency[19].

*Corresponding Author

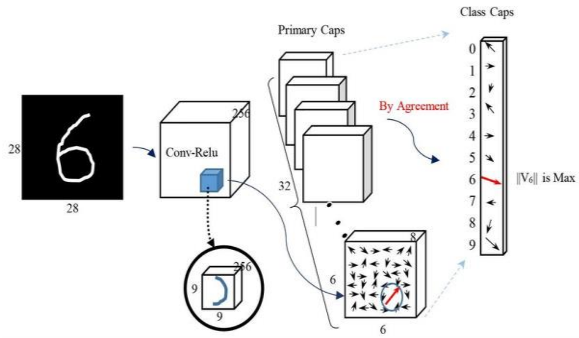


Fig. 1. CapsNet Architecture.

II. Related Work

A. EEG Feature Extraction

This study shows that the broad range of the EEG important aspect of many of the other eradication approaches for prescribe the last 12 years. Many of the essential old EEG features are used for the extraction approaches for the total focus on the frequency dimension and other related dimension information. This time is the sphere for the study about the EEG signal with the proper sequence of series[23]. Time-domain approaches are not used for the prevent. Moreover, much learning has done for human emotion like all the domain time for the better attribute. If you say about the frequency domain, the EEG signal by the relevant area uses the power of the divide-frequency approach like the alpha, beta, theta, and last is delta attribute. A crucial significant algorithm is used Fast Fourier Transform (FFT) Another opportunity we can use as an algorithm called short-time furrier Transform (STFT), and Power Spectral Density (PSD) is widely used for better understanding EEG[3]. Since the EEG signal is non-static for the people to describe the best and too much newly connecting by using the time factor, and another important thing is the frequency for the best domain access the extra information. Hilbert–Huang Transform (HHT) is one of the best learning about the EEG signal and other frequency domain. We always divided into the signal and other Intrinsic Mode Functions (IMF) by using the new trend and other than access rapid frequency data. So, the EEG feature al about the music linking subject for better understanding[18].There are many time frequencies used for the better resistant to over the noise other than the STFT based feature. First is the most frequency feature is used called (HHT) to improving the eradication and many more scale for the collapse as the EEG emotional features. So, if we say all the relevant results to first is to determine the time-frequency function to accept the most relevant result rather than the domain of the old signal for the best feature[1].We always see a better understanding of the most considerable interaction for the spatial like the time and other relevant frequency and different dimensions[9]. Moreover, the old study is all about the paint attraction to the appropriate domain. The relevant information is the best way describe the study were used to the limitation of the inequality used for the different aspect between the electrode combination an essential method is used for the calculated difference in the dominant line of the many other corresponding electrodes combination for the right and left a fraction of the scale [13]. EEG action I always used the sequence of the topologies for preventing many of the other separated multi-spectral picture for the learning about the subjective function. However, sometimes, we can see that the minor emotions used with additional spatial information [10]. An essential component of the EEG, many other directions for viewing the different Constance, established the proper circulation of the EEG electrodes (20-29) scheme of the map better frequency domain the many of the other two dimensions picture. So, a better method is under for the sequence of the image from the successive time for different EEG signals[20]. Fig.2 shows the mechanism of feature extraction used in different neural models. TABLE I shows the EEG emotion extraction of artificial neural network with different methods.

B. Emotional model

Different models are used for emotion classification, but two significant emotion models are used to understand the EEG signals. First, a necessary and essential method is used to accept a person's emotion design. All emotions are divided into six fundamental thing dislike, hearted, delight, scare, shock, sorrow. The phenomena are elaborate to hold the human emotion is ultimately showing the necessary person's emotion. Besides this, another model is used called a two-dimensional emotional model to describe the vision of the emotion of the multi-dimension[27].

The emotional element focuses on the two-dimensional plan. The first emotional step dimension represents the valence that meant by the emotional state of the person to show the range between the negative and positive emotions (Angry, Negative). The second dimension is describing the arousal. It meant that the emotional depth to feel any person. This dimension is also critical because all the simplicity and adaptability describe in this dimension. Valence-Arousal model is a beneficial classification model for better understanding[1]. Fig. 3 represents the different human emotions.

C. Emotion Classification Methods

In this paper we provide the comparison of the different methods to elaborate human emotional classification based on valance and Arousal. In below figure show the rating of the accuracy and the many other object counts as a number. In below table describes the most commonly emotional classification approaches included the K-Nearest Neighbor, that is used for the support of the vector machine represented by the SVM that is used mostly. This method is always used for the baseline method and another comparison method for help to understand this model[17].

One of the meaningful method listed to the classified by the motion of the statics and except for a new technique used in accept by learning the LSTM/RNN by using the EEG features increase periodically and dynamic[4]. Many points are used to discuss the EEG model, but there are two different points applied to the best reason for selecting the CNN and LSTM/RNN is the best part for the classification method. Another column of the table is used to show the primary classification of the different numbers of the class used in the previous study. In early research is used for the essentially divide the emotion into a subcategory [33].TABLE II shows a summary of different studies on human emotions using different neural models.

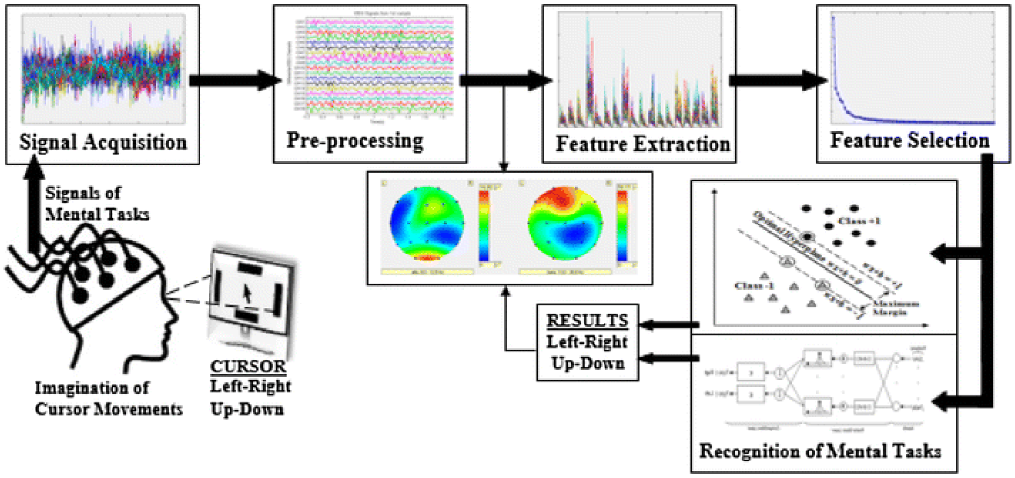


Fig. 2. EEG Feature Extraction.

TABLE. I. EEG SUMMARY FOR EMOTION EXTRACTION

Author and Study	Year	EEG Feature	Extraction Method	Dimension
Agrafioti et al.[1]	2011	Activity, Mobility, Complexity	Servick's Method	Time
Baird et al.[3]	2011	9 Sub-band of EEG (6-22 Hz)	STFT	Frequency
Candra et al.[6]	2015	Activity, Mobility, Complexity	Welch's Method	Time
Daugbjerg et al.[8]	2018	Sub-band:0, α, β, γ	FFT	Frequency
Mislove et al.[23]	2007	EEG γ band(32-100Hz)	FFT	Frequency
Orlowski et al.[25]	2010	Higher order crossing	DWT	Time
Xin et al.[16]	2017	Sub-band δ,0, α (9-14 Hz), β(2-29Hz), γ(36-40Hz)	STFT	Frequency
Thammasan et al.[29]	2016	Sub-band:0, α, β, γ, NLD, NSI	Welch's Method	Frequency
Lua et al.[22]	2011	β/ α Sub-band: β	FFT	Frequency
Kim et al.[18]	2007	HHS based feature vectors	HHS	Time and Frequency
Ghassemlooy et al.[15]	2019	Change and asymmetry in sub-band of α	Welch's Method	Frequency and Special
Crocioni et al.[7]	2007	Spectral power and difference	FFT	Frequency and Special
Bonheur et al.[5]	2019	Correlation, Coherence and phase synchronization	FFT	Frequency
Easley et al.[11]	2018	PSD, DE, DASM and RASM	STFT	Frequency
Lee et al.[19]	2014	Density Estimate	Kernel Estimation Density	Frequency
Murugappan et al.[24]	2010	Sub-band:0, α, β,	MDFDA	Frequency
Yan et al[.[32]	2014	Sum of Squared of absolute values of sub-band: 0, α, β	FFT	Frequency and Special
Skarsgard et al.[28]	2008	Fractal Dimension (FD) and Power Spectral Density (PSD)	Welch's Method	Frequency
Petrantonakis et al.[27]	2010	PSD, DE, DASM and RASM, ASM and DCAU	STFT	Frequency
Hunt et al.[16]	2012	Multi-scale Entropy	HHT	Time and Frequency
Frieden et al.[14]	2006	Frequency Features and Time-Frequency Feature	FFT	Time and Frequency

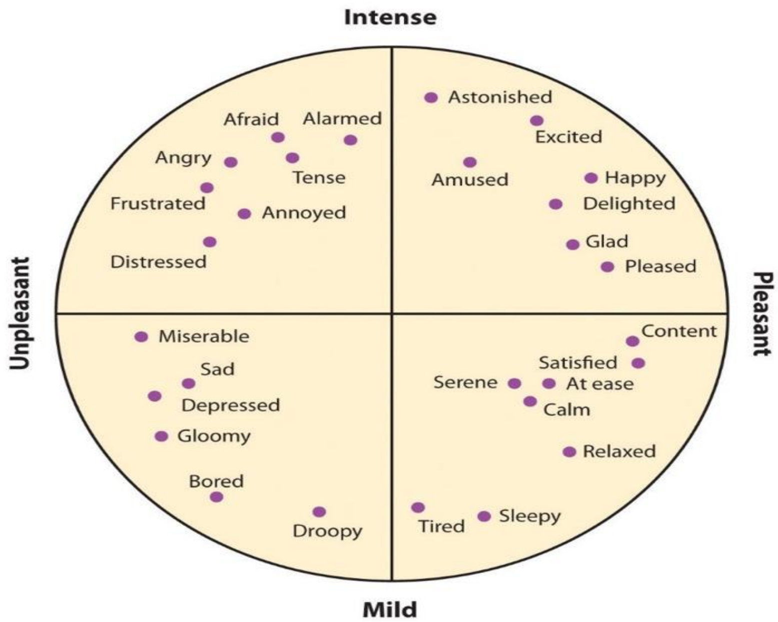


Fig. 3. Different Human Emotion.

TABLE. II. EEG SUMMARY FOR EMOTION EXTRACTION

Author and Study	Emotion Classification	Subject	Accuracy	Classification Method
Baccelli[2]	Valence and Arousal (2 class)	10	81%	SVM
Fischer[12]	Valence and Arousal (3 class)	30	66.7%	SVM
Jung[17]	Valence and Arousal (2 class, Respectively)	28	81.3%	SVM
Lua[21]	Valence (2 class)	12	71.3%	K-NN
Voellmy[30]	Valence (2 class)	9	82%	SVM, K-NN
Daugbjerg[17]	Valence and Arousal(2 class)	15	82%	SVM
Ghassemlooy[15]	Valence and Arousal (2/3 class, Respectively)	32	66.6%,66.4% (2) 53.3%,51.01% (3)	Bayes Neural Network
Oh[25]	Valence and Arousal (2 class, Respectively)	32	74.12%	C-RNN
Voellmy[30]	Valence and Arousal(4 class)	32	75.21%	CLRNN

D. EEG MFI Progressions

The universal scheme recognizes the 12-22 is ides for the best design of identity apply for the necessary edge of an amount to electrodes the situation for the EEG testing. This scheme is based on the communication between the best edge of the wires and the other underlying area for the intelligence. The difference shown between the (13 to 23) assigns the essential fact to calculate the real distance between the adjacent electrodes are either the 12% t or 22% for the all front-back to the left-right distance of the scalp[22]. An international plan in view roundabout (10-20) system and another generalized squire of the matrix used for it. In the EEG electrodes, point to describe the best test used for the DEAP datasets. We can see that the DEAP dataset is used in the square matrix (N x N) [4]. In the following equation, N, we used for the extreme point of the number exist between the vertical and horizontal. In looking for the DEAP dataset, Where N is equal to the 10. So, the basic matrixes are used to fill the EEG frequency that represents a side ofthe right of the picture that shows clearly. Thus, the line is gray looks like the triangle that is mention before the midpoint of the square matrixes describes the mansion. Many other color combinations show mostly red point shows the electrodes comparable to the red circle used for the universal (12-22) edge, and the gray color point is inserted of other the full matrixes. The essential cost of red color is showing comparable frequency-feature (PSD) that is used for the EEG electrodes. If we see in the pointed gray that is an addition to the red point neighboring that [31]. Fig 4 represents the EEG MFI node construction.

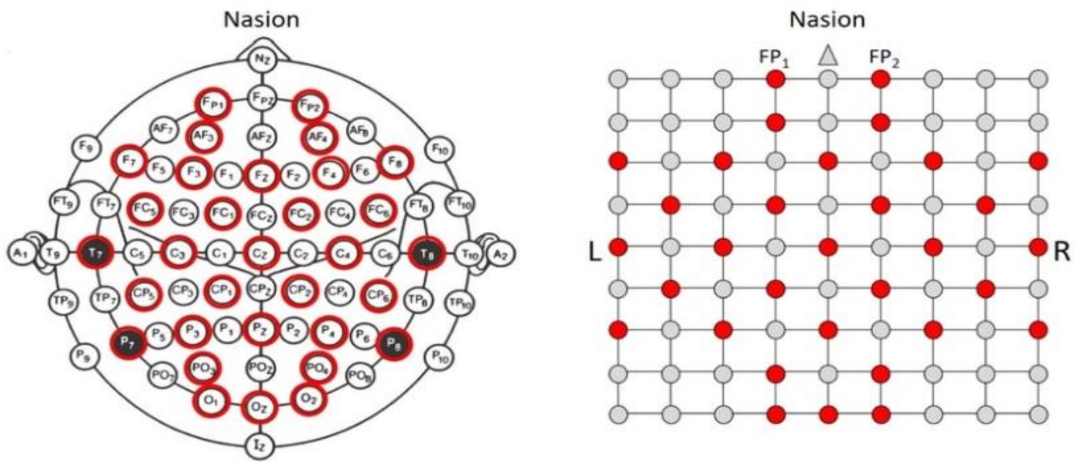


Fig. 4. Sequences of EEG MFI Construction.

III. Caps-Net Neural Network

CapsNet neural network is used for the neural vector other than a neural point for the old neural network track of network model changing routing techniques. It is mostly studying the feature of the training data fully classify the details[2]. The old CNN neural network is mainly used local classified feature for the speech, video and image fields, moreover CNN demands to complete a copy of the feature detector. If we draw a design for the spatial information that can diminish the efficiency of one structure. There is not a limited efficiency model. Still, sometimes linear and distinct data not use for the EEG signals, CNN neural network not adequately described for the EEG signal features and other classify efficiency changing routing technique in Caps Net network that can create a fact between lowest level and heights level point. Which can be used as an intelligent model for the better relationship between the local and all other models. That is automatically applying the network to sense for the new climate[25]. Fig5 represents the CapsNet neural network model related to human emotion signals.

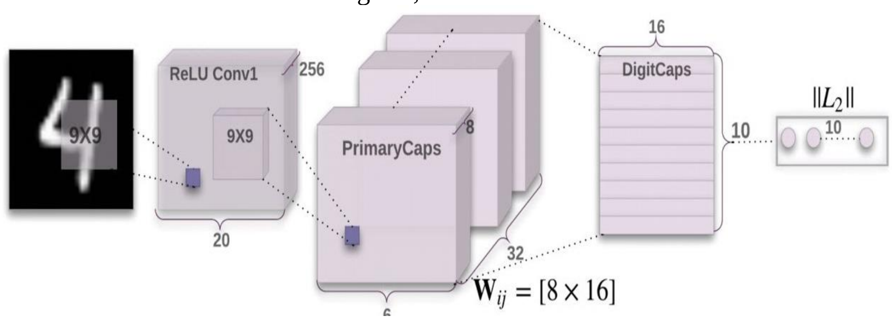


Fig. 5. CapsNet Neural Network.

IV. Deep Learning based in CNN

Deep learning based on the access for the two different picture qualification techniques, first is the CNN and another is the Cap-Net. Both are used for the general classification task. We can concert the AD EEG signals into the 2D picture by using the STFT algorithm. STFT algorithm produces the 2D time-frequency sphere picture from any channel like electrodes [25]. For example, if we want to use the EEG signal from the 3 or 4 electrodes, then 3 channels are using the 2D image data look like RGB picture generating and use the testing and training of the classifier. When we calculated the suppress function that can act like the non-linear active function of CNN to develop and used the training and the testing of the classifier. Many of the other details about the generation in 2D STFT picture is used for the calculated 3 electrodes with the examiner rate is 255 Hz is described before. After using the EEG signal to accept the notch-filtered that is used for the separation according to the best action performs before. We can use the data of the array size 4 x 2 x 400. If we apply the algorithm STFT, we can draw the 2D data array size 79 x 16 which is used for the X-axis and y-axis that signed to the frequency and time respectively[6].In EEG, signals can be divided into different band frequency like Alpha (9-18Hz), Beta (18-33Hz), Gamma (>34Hz), Theta (5-8Hz), Delta (<5Hz), and Mu (9-13Hz). They all depend on the desire and operation of the study for the specific range that can be calculated by the signals (e.g., mu and other Beta Bands) which is selective. For motor symbolism allocation tasks like the frequency band, 6-32Hz are commonly used. In this paper, we study to choose the frequency range calculates the 9-31Hz for the quotation of the STFT spectrogram picture. So, in the end, we can see 3 channel 14 x 15 image in capsule networks that can be used as a testing and training thoroughly[7]. Clarify the network design to propose the Caps Net-based path. At the same original Caps-Net paper takes out the 2D picture as the starting input. We can see in our architecture layer (kernel size: 8x8 with 2 strides 2), another dropout layer before we can use the Caps layer, the result shows the 18 dimensions as per class (Left-Hand as compare to the Right-Hand). The training of the Caps Net is also complete. The versatile routing by agreement exists between the capsule as detailed before. The learning ability that can be feasible by the Caps Net on the EEG domain also explained in the previous section[30].Fig 6 shows the deep learning CNN model.

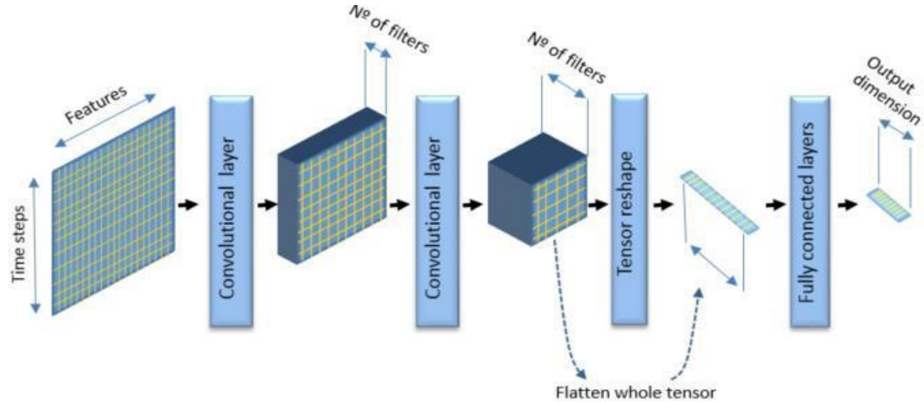


Fig. 6. Deep Learning in CNN.

V. Convolutional Neural Networks

The necessary input used for MFI is to calculate the size 250 x 198 pixels, and it consists of three different colors of channel. Many of the set numbers that can be used for the twist filtering as 28 for the first convolutional layer to be use the excerpt for the 29 different types of correlation information name as the 28 various features, In that time we extract the different ratio dimensional layer[28]. There are many sizes of the field used like 5 x 6 pixels and 8 x 8 pixels commonly. Usually there are different sizes used like the stalk 3,6 and 11 pixels because it won’t have any overlap and strides. The primary activation action is used like ReLU[7]. The firstly ordinary layers used in max-pooling and this pooling size is 3 x 3, these are the strides. A second regular layer uses 10 different filtering with their size 4 x 4 without any overlap between strides. Further integrate information is always helpful for the exclusive range for the previous feature's unit is used for connecting. A smash operation has occupied the change of the final feature into the different spatial functions of the vector. The basic structure of CNN is explained earlier. The solid layer changes the final feature into the one-spatial feature vector. Looking is this layer; we can easily set the output 1/11 of the soloed layer. Furthermore RNN output layer SoftMax is used as the real activate function for the exact output size prior to other 5 corresponding emotion state[21].

Table III shows the specifications of different CNN layers.

TABLE. III. Convolutional Neural Networks

Layers	**Specification**
Convolution (Conv) 1a	64(Filter:3x3x3 and stride: 1x1x1)
Pooling 1	1x2x1
Conv.2a	128(Filter:3x3x3 and stride: 1x1x1)
Pooling 2	2x2x2
Conv.3a	256(Filter:3x3x3 and stride: 1x1x1)
Conv.3b	256(Filter:3x3x3 and stride: 1x1x1)
Pooling 3	2x2x2
Conv.4a	512(Filter:3x3x3 and stride: 1x1x1)
Conv.4b	512(Filter:3x3x3 and stride: 1x1x1)
Conv.5b	512(Filter:3x3x3 and stride: 1x1x1)
Pooling 5	2x2x2
Fully Connected (FC)	4096
FC	4096
Soft Max	Output

VI. LSTM Recurrent Neural Network

Many of the in-depth analysis of simulation emotion change up the 70s. This emotion is calculating by the different subjects often fare the important moving part to all videos. Moreover, we want to describe a structure for the text information for many long distances as per sequence. In the previous look the RNN is the better sequential structure[29]. Simple RNN challenges see or in gradient vanish, or others see in the collapse in the back reproduction. So, the LSTM unit is used for the accept the easy group for old RNN, and the LSTM unit is the combination of the entering mechanism for the better structure[12]. So, data timing is most useful thing to manage and carry easily for the long during for best computation, the doorway is used for the necessary instruction and another individual -loop mechanism. That allows the gradient to move very long durations[30].A standard design that is used in the LSTM unit to describe in the below figure. If we compare two different structures of the neural network units. There are two units illustrated in the below picture and comparison between the recurrent neural network and LSTM [32]. In the picture, we can see that RNN is used for the content of the evolution of the output that given input. The other LSTM is used for the content the three basic gate structure like (input gate, forget gate, and output gate) that is used to observe what kind of the information is move from the previous step and it should be erased the information at the current time for added the main data flow [16].Fig 7 shows the functioning of LSTM Recurrent Neural Network.

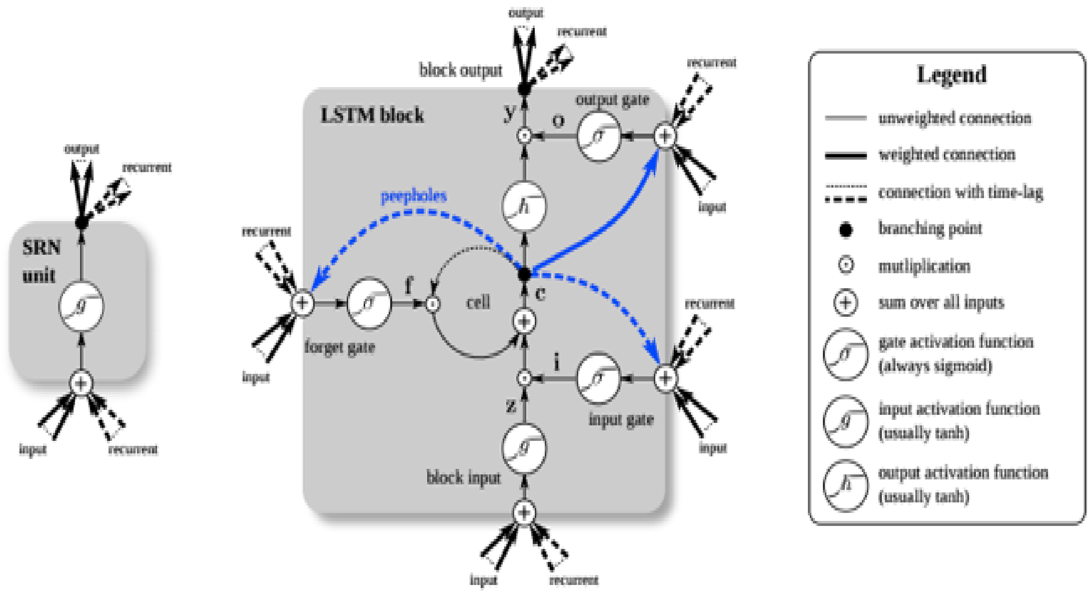


Fig. 7. LSTM Recurrent Neural Network.

VII. Results and Discussion

A. Dataset and Preprocessing

We can calculate the actual work expected into the Caps Net using basic on the best path on the BCI dataset. We can organize the massive experiment by using the BCI competition VI 2c datasets. That dataset can be a subset of EEG signals which can calculate the essential 10 subjects between two different class motor symbolism function. For example (left side and right side).There are 4 primary bipolar channels used like (C2, C5, and C6) for the EEG that can be recorded with different frequency of 260Hz. Also, there are many signals existing between 0.8 and 123Hz used for the filter bandpass bit notch filter also apply for 54Hz frequency can handle almost 6 data set consist in their session. The first four sessions are used for the practice, and other is used for the testing. The average of the training data for each subject is 450 trails to count (200+200+50) where test data just for the 300(150+150). By using the experiment to count the effectiveness of the dataset. A classifier is used for the train and test subject by subject experiment can be done according to the guidelines of the data set for the relevant classifier is used for the test and training motor-imagery activation. There are many semi motors to capture the μ and β bands. Furthermore, an excellent frequency band is used. For example (9Hz ~ 90Hz) are mostly chosen by the feature of extraction. A good understanding of the EEG division used between the 0.7 to 3.1 that can be greeted by a better classification result. There are many experiments used for the EEG original signals from the dataset that have filtered the band pass exist between the 9~33Hz and just 3 EEG segment (0.7~3.5s) is used for processing the features. A significant network can be used to analyze the complete Brain Decode framework. A lot of features can be used for the Processing of EEG. The study can be controlled by the desktop PC computers will be used called X GPU. Fig 8 shows different datasets of CapsNet.

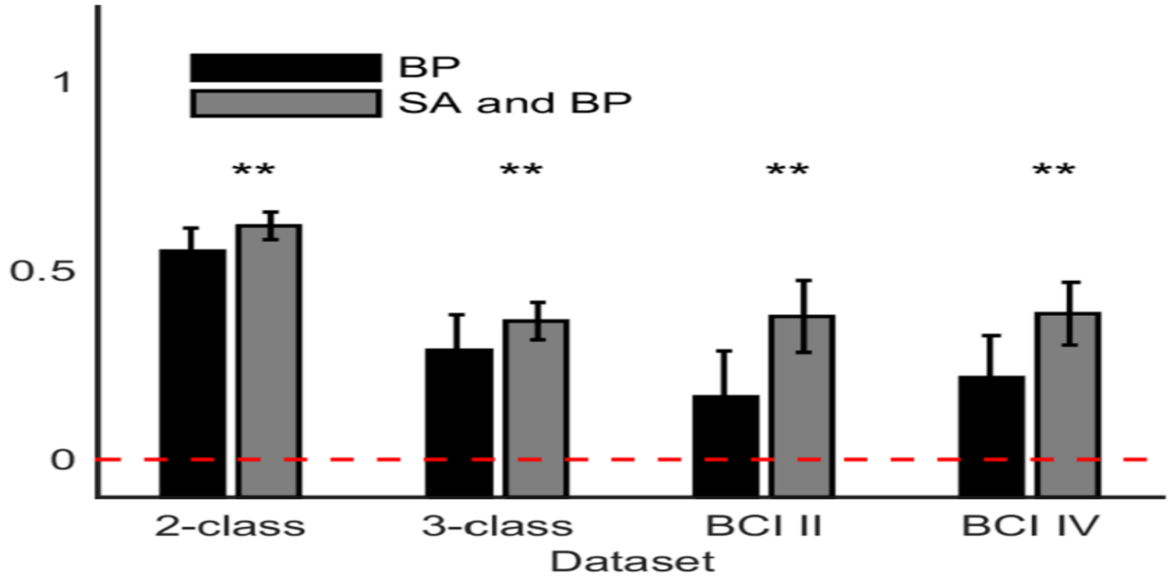


Fig. 8. BCI Dataset of CapsNet.

B. Performance Evaluation

Firstly, we can calculate the feasibility for apply the Caps Net on the EEG motor symbolism for the different analysis that can be used for the original design to better picture and point classification. The Caps Net is trained for every case with the 35 initial capsule that is the best pathway used 9 and 19 detention for MI caps with 2 output right and left. The Caps Net is trained for the best EEG sphere with the STFT picture. We can see that the test and train are decreased fast within 15 times and gradually concentrate for 3 times. The best performance is used many terms to reduce the best efficiency express 3 minutes classification rates like the many numbers of the trails out of all paths. The Caps Net attain 86% of different classification efficiency. When we use many numbers of routing repetition is set to 6. It does not found any vital effects that can be used for the number of repetitions to classify the accurate network. We can check the initial results of the process after repair for best performance. Including the best analysis, the learning ability for the Caps Net for the EEG signal domain. If we can make a comparison between the different classification of the accuracy used for the expected Caps Net based approach with the old CNN based approach, in comparison of different Caps Net apply for the best suggestion by the 2 different involutions and pooling layer for the different decryption motor for the symbolism signal using SFTF picture. If we test and train, the input of the CNN approaches in the same STFT picture using the 6 channels 16x16 2D vectors. In this analysis, the primary classification can be accurate and best average calculation scores for the 15 different subjects evaluates. TABLE 4 shows the comparison between CNN and CapsNet. Fig. 9 shows graphic view of this comparison.

TABLE. IV. COMPARISON BETWEEN THE CNN AND CAPSNET

Subject	**CNN**	**CapsNet (Planned)**
01	71.72%	77.01%
02	51.15%	59.54%
03	54.18%	49.12%
04	90.38%	91.23%
05	80.18%	84.25%
06	74.01%	92.19%
07	71.25%	71.24%
08	84.58%	88.21%
09	83.41%	80.31%
Average	73.43%	70.01%

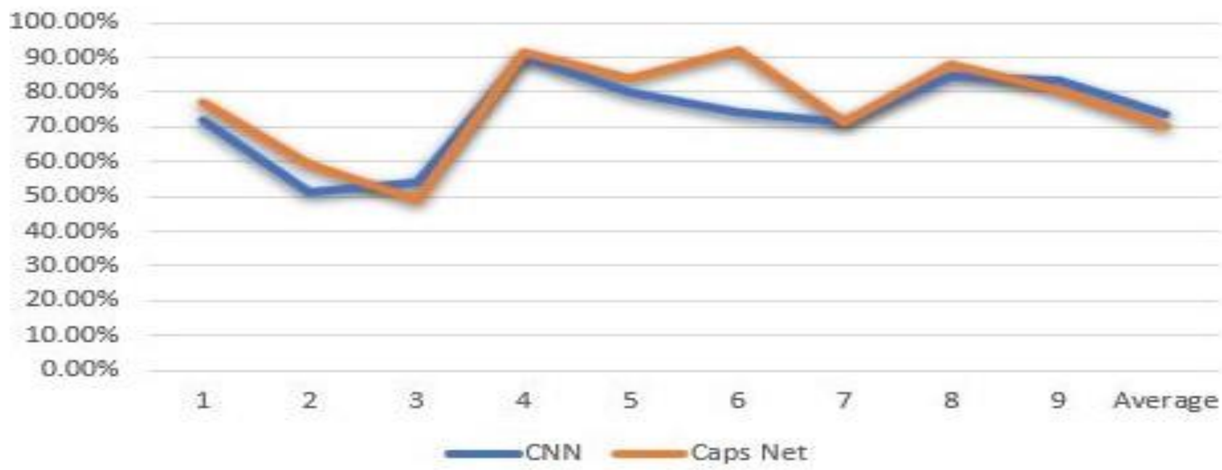


Fig. 9. Comparison between the CNN and CapsNet.

VIII. Conclusions

In this paper, firstly classify the Caps-Net neural network and observe the primary emotion of the EEG signals. The total result must be shown that the Caps-Net achieves precise information collection rather than the old neural network. There are many useful emotional classifications of the EEG signal. At the same time, the basic design and framework are adopted to the Caps Net. Calculate the better performance and feasibility that will describe the different methods for many experiments on BCI completion was attended. Another experimental result conducted previously for different methods of using CNN based two class motor imagery tests. Moreover, the improvement of the Caps-Net based approach uses EEG decryption frequency features (PSD) are used to exact the EEG channel and made the two-dimensional design for construct the EEG MFI. So, the necessary signals of EEG MI are building for the apparent sequence of the raw signals. Hybrid deep neural networks are connected to the primary convolution of the Neural Network, and many of the below-term memory recurrent neural networking hybrid design CNN is used for the sense of temporary picture EEG MFI sequence of the pattern. To clarify, the LSM/RNN is used for human emotions. If we look at the future side, there are many deep searches employed for classify the underlying protocol of the EEG signal.

References

[1] F.Agrafioti, D. Hatzinakos, and A.K. Anderson "ECG pattern analysis for emotion detection", IEEE T AFFECT COMPUT., vol. 3, no.1, pp. 102-115, 2011.

[2] F. Baccelli and B. Błaszczyszyn, "Stochastic geometry and wireless networks: Volume II Applications," Foundations and Trends® in Networking., vol. 4, no.1–2, pp. 1-312, 2010.

[3] R. Baird, G. Eeson, A.Safavi, P. Puligandla, J. M. Laberge, E. D.Skarsgard and C. P. S. Netorks, "Institutional practice and outcome variation in the management of congenital diaphragmatic hernia and gastroschisis in Canada: a report from the Canadian Pediatric Surgery Network," J. Pediatr. Surg., vol. 46, no.5, pp. 801-807, 2011.

[4] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All your contacts are belong to us: automated identity theft attacks on social networks, ". Proc. 18th Int. Conf. W.W.W, ACM., 2009.

[5] S.Bonheur, D. Štern,C. Payer, M. Pienn, H. Olschewski and M. Urschler, "Matwo-CapsNet: "A Multi-label Semantic Segmentation Capsules Network," Int. Conf. Med. Image. Comput. Comput. Assist. Interv., Springer,2019.

[6] H. Candra, M. Yuwono, R.Chai, A, Handojoseno, I. Elamvazuthi, H.T. Nguyen and S.Su, "Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine," 2015 37thAnnual int. conf.IEEE Eng. Med. Biol. Soc., pp.7250-7253 .

[7] P. Crocioni, "Net neutrality in Europe: Desperately seeking a market failure," Telecommun. Policy., vol. 35, no. 1, pp.1-11,2011.

[8] C. Daugbjerg, "Policy networks under pressure: pollution control, policy reform and the power of farmers, Routledge," 2018.

[9] A. Dhamdhere and C. Dovrolis, "Can ISPs be profitable without violating "network neutralit"?," Proceedings of the 3rd international workshop on Economics of networked systems, ACM.,pp.13-18,2008.

[10] M. Dischinger, A. Haeberlen, K.P. Gummadi and S. Saroiu, "Characterizing residential broadband networks," (2007).. Internet measurement conference, vol. 7, pp. 43-56.

[11] R. Easley and F. H. Guo, "Research Commentary—From Net Neutrality to Data Neutrality: A Techno-Economic Framework and Research Agenda." Info. Syst. Res., vol. 29, no.2, pp. 253-272, 2018.

[12] A.Fischer, J. F. Botero, M.T. Beck, H. De Meer and X. Hesselbach, "Virtual network embedding A survey," IEEE Commun. Surv. Tut., vol. 15, no. 4, pp.1888-1906, 2013.

[13] T.Fischer, and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," Eur. J. Oper. Res., vol. 270, no.2, pp. 654-669, 2018.

[14] R.Frieden, "Network Neutrality or Bias-Handicapping the Odds for a Tiered and Branded Internet," Hastings Comm. & Ent. LJ ., vol. 29, pp.171,2006.

[15] Z. Ghassemlooy, W. Popoola and S. Rajbhandari, "Optical wireless communications: system and channel modelling with Matlab®, CRC press," 2019.

[16] K.J. Hunt, G. R. Irwin and K.Warwick,"Neural network engineering in dynamic control systems," Springer Science & Business Media, 2012.

[17] C. Y.Jung, H. Y. Hwang, D. K. Sung and G. U. Hwang , "Enhanced Markov chain model and throughput analysis of the slotted CSMA/CA for IEEE 802.15. 4 under unsaturated traffic conditions." IEEE Transactions on Vehicular Technology, vol.58, no. 1, pp. 473-478, 2008.

[18] S.Kim, S. Pakzad, D.Culler, J. Demmel, G. Fenves, S. Glaser, and M. Turon,"Health monitoring of civil infrastructures using wireless sensor networks," Proceedings of the 6th international conference on Information processing in sensor networks, ACM, pp. 254-263, 2007.

[19] Y.Y. Lee and S. Hsieh, "Classifying different emotional states by means of EEG-based functional connectivity patterns," PloS one , vol.9, no. 4, p.e95415, 2014.

[20] Y.Li, W. Zheng, Z. Cui, Y. Zong and S. Ge, "EEG emotion recognition based on graph regularized sparse linear regression." Neural Process Lett., vol. 49, no. 2, pp. 555-571, 2019.

[21] R.Luaand K. C. Yow, "Mitigating ddos attacks with transparent and intelligent fast-flux swarm network," IEEE Network, vol. 25, no.4, pp.28-33, 2011.

[22] A.Mert, and A. Akan, "Emotion recognition from EEG signals by using multivariate empirical mode decomposition," Pattern Anal Appl., vol. 21, no.1, pp.81-89, 2018.

[23] A.Mislove, M. Marcon, K. P.Gummadi, P.Druschel and B.Bhattacharjee,"Measurement and analysis of online social networks," Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, ACM, pp. 29-42, 2007.

[24] H.Oh, , G. Labianca and G. H. Chung , "A multilevel model of group social capital," Academy of management review , vol. 31, no.3, pp. 569-582, 2006.

[25] S.Orlowski and R. Wessäly, "SNDlib 1.0—Survivable network design library," Networks: An International Journal, vol. 55, no.3, pp.276-286, 2010.

[26] P. C.Petrantonakis, and L. J. Hadjileontiadis "Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis," IEEE Transactions on affective computing, vol. 1, no. 2, pp. 81-97, 2010.

[27] D.Skarsgard, J. Claydon, S. Bouchard, P. C. Kim, S. K. Lee, J. M. Laberge, D. McMillan, P. von. Dadelszen and N. Yanchar,"Canadian Pediatric Surgical Network: a population-based pediatric surgery network and database for analyzing surgical birth defects. The first 100 cases of gastroschisis," J. ped. surg. vol. 43, no. 1, pp. 30-34, 2008.

[28] N.Thammasan, K. Moriyama, K. I. Fukui and M. Numao "Continuous music-emotion recognition based on electroencephalogram." IEICE TRANSACTIONS on Information and Systems, vol. 99, no.4, pp.1234-1241, 2016.

[29] A.Voellmy and H. Kim, et al. Procera: a language for high-level reactive network control. Proceedings of the first workshop on Hot topics in software defined networks, ACM, 2012.

[30] L.Xin, Q. Xiaoying, S. Xiaoqi, X. Jiali, F. Mengdi and K. Jiannan,"An improved multi-scale entropy algorithm in emotion EEG features extraction." J. Med. Imag Heal Info. vol. 7, no. 2, pp.436-439, 2017.

[31] J.Yan and W. Zheng, "Integrating facial expression and body gesture in videos for emotion recognition," IEICE TRANSACTIONS on Information and Systems , vol. 97, no. pp. 610-613, 2014.

[32] X.Yu and Q. Gu, Citation prediction in heterogeneous bibliographic networks. Proceedings of the 2012 SIAM International Conference on Data Mining, SIAM, 2012.

[33] W.L.Zheng and B.-N. Dong, Multimodal emotion recognition using EEG and eye tracking data. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2014.

Behavior of Learning Rules in Hopfield Neural Network for Odia Script

Ramesh Chandra Sahoo[1]
Research Scholar, Dept. of Computer Science and Applications, Utkal University, Bhubaneswar, India

Sateesh Kumar Pradhan[2]
Dept. of Computer Science & Applications
Utkal University, Bhubaneswar, India

***Abstract*—Automatic character recognition is one of the challenging fields in pattern recognition especially for handwritten Odia characters as many of these characters are similar and rounded in shape. In this paper, a comparative performance analysis of Hopfield neural network for storing and recalling of handwritten and printed Odia characters with three different learning rules such as Hebbian, Pseudo-inverse and Storkey learning rule has been presented. An experimental exploration of these three learning rules in Hopfield network has been performed in two different ways to measure the performance of the network to corrupted patterns. In the first experimental work, an attempt has been proposed to demonstrate the performance of storing and recalling of Odia characters (vowels and consonants) in image form of size 30 X 30 on Hopfield network with different noise percentages. At the same time, the performance of recognition accuracy has been observed by partitioning the dataset into training and a different testing dataset with *k*-fold cross-validation method in the second experimental attempt. The simulation results obtained in this study express the comparative performance of the network for recalling of stored patterns and recognizing a new set of testing patterns with various noise percentages for different learning rules.**

Keywords*—Hopfield network; Odia script; Hebbian; pseudo-inverse; Storkey; NIT dataset

I. Introduction

Odia language is one the oldest and official language of Odisha state in the Indian subcontinent used by more than 50 million people. Modern Odia language consists of 47 different characters out of which 11 are vowels and rest are consonants. Automatic recognition of these Odia characters is particularly difficult because many of these characters are similar looking rounded in shape. Many works have been done so far in the field of OCR (Optical Character Recognition) for various Indian languages but very little research has been done for Odia language. In this study we focused on 11 Odia vowels and 36 basic consonants for storing into and recalling from Hopfield network with different learning rules especially when the patterns gets distorted. We add explicit noise into the input patterns from 10% to 50% while recalling them and compare their results for analysis. The main idea for choosing Hopfield network is that it can be a multiple point attractors for high dimensional space and due to the dynamics of network that guaranteed to convergence to local minima. Hopfield network is an auto associative memory network that reproduces its input pattern as an output even if the input pattern is distorted or partially feed into the network. It a fully connected special type of recurrent network excluding self connection with binary or bipolar inputs. The network returns a stored best matching pattern when an input pattern is presented to the network. The experimental simulation have been carried out in MATLAB2014a environment and the performance parameters such as recalling stored patterns and recognition accuracy of new patterns with noisy version is validate and compared with other techniques in the literature.

II. Related Work

Many articles have been published for various Indian languages like Telgu [1], Tamil [2], Gujarati [3], Kannada [4], Bangla [7] and Gurmukhi [5]. Whereas for Odia character recognition, there are few articles available for study. Development of a novel OCR system for Odia character is difficult and challenging task because most of these characters are identical and roundish shape. In 2002 Chaudhuri et al [6] suggested a system for printed odia script with 96.3% accuracy. Mohanti [8] proposed a system by using Kohonen network to recognize Odia alphabets. Roy et al. [9] proposed a system for Odia handwritten numerals with chain code histogram with a accuracy of 94.8%. Bhowmik et al. [10] suggested a system for Odia handwritten digits with approximately 93% accuracy by using hidden markov as a classifier. Sarangi et al. [11] proposed a classifier by using Hopfield neural network for Odia numerals. Panda et al. [12] proposed a single layer perception for Odia handwritten digits recognition by using gradient and curvature feature extraction methods with an accuracy of 85%. Das et al.[13] presented a hybrid method for Odia digit classification using Kirsch operator. Meher et al.[14] proposed a method for Odia character recognition in the presence of vowel modifiers of complex Odia characters with an accuracy of 91.24%. Kumar et al.[15] proposed an Ant-Miner algorithm for Odia character recognition with an accuracy of 90%. Mishra et al.[16] proposed a fuzzy based algorithm using HMM classifier for handwritten characters with an accuracy of 96.3%. Pujari et al.[17] compares different algorithms proposed so far for Odia character recognition. Chen L –C et al.[18] proposed a modified Hopfield network for character recognition system by using a single error correction-single error detection method to improve the learning in the network. M.B. Sukhaswami et al.[19] proposed a method for Telgu character recognition using Hopfield network and then modified it to Multiple Neural Network Associative Memory (MNNAM) to overcome the storage capacity limitation of Hopfield network. Prateek Malik et al.[20] proposed handwritten character recognition using Hopfield network and wavelet transforms with various noise percentage. Kalyan S Dash et al. [25], Proposed a hybrid feature based model for Odia numerals with ISI Kolkata database. With hybrid feature and DLQDF classifier they claimed a 98.50% of recognition accuracy and with MQDF classifier they achieved 98.40% of accuracy. Kalyan S Dash et al. [26] presented a review report on various preprocessing, segmentation and feature extraction techniques with several classifiers for Odia character recognition. They also created a standard alphanumeric Odia handwritten character database. The comparisons presented there are being taken into consideration for this work. Indugu Rushiraj et al. [27] Achieved 87.60% accuracy for Odia consonants classification by using shadow, centroid and distance based features with weighted Euclidean distance method for classification. Smruti Rekha Panda and Jogeswar Tripathy [28], proposed a template matching with Unicode mapping for Odia offline character recognition with a accuracy of 97% for basic characters and numerals. Kalyan S Dash et al. [29], proposed a novel feature extraction method called BESAC: binary external symmetry axis constellation with a Boolean matching character recognition technique for Odia and Bangla Numerals and characters. They achieved 99.35% accuracy with ISI Kolkata Odia Numerals, 98.90% for IITBBS Odia numerals, 99.48% for Bangla Numerals and an accuracy of 95.01% for Odia Characters. Subhashree Satapathy et al. [30], proposed a deep neural Autoencoder for dimensionality reduction as well as classification of Odia digits. They achieved a 97.63% of accuracy on 37 training set and different test accuracies obtained as 94.69% on 15 training sets, 97.1% on 20 training sets and 97.4% accuracy on 30 training sets. In Table I, a detail study on Odia characters and numerals recognition is shown with their claimed accuracy.

III. Odia Character Set and Chanlleges

Odia language is the mother tongue of the Indian state of Odisha and the script originally came from Brahmi script. The way of writing Odia language is unique and different from other regional languages in India. Modern Odia script consists of 11 vowels, 36 consonants and ten digits. Apart from these basic Odia character set, the Odia ligatures may be formed either merging a vowel diacritic with the consonant or by clustering two or more consonants. There are nearly 116 composite characters also known as Juktakhyaras. A printed version of Odia characters and numerals are shown in Fig. 1, 2 and 3 for vowels, consonants and numerals respectively. Most of the Odia characters are roundish in shape and very similar to each other. Recognition of Odia characters are more challenging pattern recognition task as i) many roundish and similar shape, ii) there is no reference line as in other Indian regional languages like Hindi and Bangla, iii) unavailability of more number of standard databases for both printed and handwritten Odia characters. Few samples of Odia characters with almost similar in shape are shown in Fig. 4.

IV. Hopfield Neural Network

Hopfield network [21] is merely the best known auto-associator neural network that acts as content addressable memory. It is a fully connected network with symmetric weight where no neuron is connected to itself. The neurons of this Hopfield network are updated asynchronously and in parallel and this type of networks guaranteed to converge a closest learnt pattern. The weight matrix of N neurons Hopfield network is given by a $N \times N$ matrix W with symmetric weights and the diagonal is set to zero. The state vector S at any point of time of the Hopfield network is given as:

$$S = [S_1, S_2, \ldots, S_N] \tag{1}$$

Where S_i is the current output state of unit i and the local net input to i^{th} unit is given by

$$h_i = \sum_{j \neq i} w_{ij} S_j, \tag{2}$$

Where w_{ij} is the weight of connection j to i unit and the next state of the of a unit is given by

$$S' = \begin{cases} +1 & if\ h_i \geq \theta \\ -1 & if\ h_i < \theta' \end{cases} \tag{3}$$

Where θ is the threshold at unit i.

V. Learning Rules

Learning rules in Hopfield network is basically finds the set of connection weights which allow the network to produce desired response when a pattern is submitted to the network. To compare the performance of Hopfield network with Odia vowel characters, three learning rules are used in this study. These learning rules are presented in the following section.

A. Hebbian Learning Rule [22]

Hebbian rule states that when one neuron stimulates another neuron from two connected neurons in the network for firing then the connection weight between these two cells is strengthened otherwise it is weakened. In other words we can say when a weight contributes to firing a neuron, the weight is increased. The Hebbian rule for updating the weights is presented below:

- Initialize all weights to zero

$$w_{ij} = 0, for\ i = \{1,2, \ldots, n\,; j = 1,2, \ldots, m\}$$

- For each input training(I)-target output(O) vector pair {I:O}, set the following activation as

$$x_i = I_i, for\ i = 1,2, .. n \ \&\ y_j = O_j, for\ j = 1,2, .. m$$

- Adjust the weight for $\{i = 1,2, .. n \ \&\ j = 1,2, .. m\}$ as follows

$$w_{ij}^{new} = w_{ij}^{old} + x_i y_j \tag{4}$$

- Now set the activation of output unit as follows

$$y_o = \begin{cases} +1 & if\ h > 0 \\ 0 & if\ h = 0 \\ -1 & if\ h < 0 \end{cases} \tag{5}$$

$$where\ h = \sum_i x_i w_{ij}$$

To store(train) K patterns $\{y_1, y_2, \ldots, y_k\}$ with a N nodes Hopfield network performs the following equation by Hebbian learning rule.

For training the two associated patterns in each pair are the same which means $x_k = y_k$ with both have same dimension i.e. $m = n$. Then the weight matrix is obtained for K patterns as the sum of their outer products as:

$$W_{n\times n} = \frac{1}{N}\sum_{k=1}^{K} Y_k X_k^T = \frac{1}{N}\sum_{k=1}^{K} Y_k Y_k^T \quad (6)$$

And in the matrix form it is as follows:

$$W_{n\times n} = \frac{1}{N}\sum_{k=1}^{K}\begin{bmatrix} y_1^k \\ y_2^k \\ \vdots \\ y_N^k \end{bmatrix}[y_1^k, y_2^k \cdots y_N^k]$$

And the weight between connecting node i to j is

$$w_{ij} = \frac{1}{N}\sum_{k=1}^{K} y_i^k\, y_j^k = w_{ji} \text{ and } w_{ii} = 0\ \forall\ i \quad (7)$$

Hebbian learning rule is incremental and local which means the previous matrix can reused when we add a new pattern to the memory that to be learn.

B. *Pseudo-Inverse Learning Rule [23]*

Pseudo-inverse learning rule uses the pseudoinverse of the pattern matrix while Hebbian learning rule uses the pattern correlation pattern matrix. When pattern vectors are not orthogonal this learning method is more efficient and it is neither local nor incremental which means a new pattern cannot incrementally added to the network and update does not depend on either side of the connection as it calculate the inverse of the weight matrix. The procedure to apply pseudo-inverse learning rule to obtain the weight matrix of Hopfield network is explained as follows.

- Let input patterns X of vectors $\{x_1, x_2, \dots, x_n\}$ and target pattern Y of vector $\{y_1, y_2, \dots, y_n\}$ then we can write:

$$\mathrm{WX} = \mathrm{Y} \quad (8)$$

$where\ W\ is\ the\ weight\ matrix$

- If the matrix X has an inverse, then we can rewrite:

$$\mathrm{W} = \mathrm{XY}^{-1} \quad (9)$$

- Again if X is not a square matrix then no exact inverse will exist, but it has been shown that it will minimize to $\mathrm{F(W)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} ||\mathrm{y_i} - \mathrm{Wx_i}||^2$ by using the pseudo-inverse matrix and is as follows:

$$\mathrm{W} = \mathrm{YX}^{+} \quad (10)$$

Where X^{+} is the Moore Penrose pseudoinverse, then the pseudoinverse of real matrix X is the unique matrix that satisfies:

$$\mathrm{XX^{+}X} = \mathrm{X}, \quad \mathrm{X^{+}XX^{+}} = \mathrm{X^{+}},$$
$$\mathrm{X^{+}X} = (\mathrm{X^{+}X})^{\mathrm{T}} \text{ and } \mathrm{XX^{+}} = (\mathrm{XX^{+}})^{\mathrm{T}}$$

- So the pseudo-inverse weight matrix can be calculated as :

$$\mathrm{W_{pinv}} = \mathrm{W^T} * (\mathrm{W} * \mathrm{W^T})^{-1} \quad (11)$$

$where\ W^T\ is\ the\ transpose\ of\ weight\ matrix$

- Hence the Pseudo-inverse learning rule is given by:

$$w_{ij}^{l} = \frac{1}{N}\sum_{l=1}^{K}\sum_{m=1}^{K} y_i^l (Q^{-1}) y_j^m \quad (12)$$

$$where\ Q = \frac{1}{N}\sum_{i=1}^{n} y_k^l y_k^m$$

Pseudo-inverse rule works well with linearly independent patterns and stores up to N patterns in an N unit network. This rule finds a set of orthogonal vectors and calculate output weight matrix by pseudoinverse solution.

C. *Storkey Learning Rule [24]*

Amos Storkey in 1997 proposed a new learning rule named as Storkey learning rule is also both local and incremental and provides greater storage capacity ($\frac{N}{2*sqrt(lnN)}$) than Hebbian learning rule ($\frac{N}{2lnN}$). It is local because it only considers neurons at either side and incremental as new pattern can be added to the network without knowledge of old pattern that have been also used for training. The learning rule mathematically defined as:

$$w_{ij}^{0} = 0\ for\ all\ i, j \quad (13)$$

And

$$w_{ij}^{p} = w_{ij}^{p-1} + \frac{1}{N}\left(y_i^p y_j^p - \frac{1}{n} y_i^p h_{ji}^p - \frac{1}{n} h_{ij}^p y_j^p \right) \quad (14)$$

Where w_{ij}^p is the weight between i and j after p^{th} pattern has been learnt and y^p is the new learning pattern.

$$h_{ij}^{p} = \sum_{k=1\ k\neq i,j}^{n} w_{ik}^{p-1} y_k^p\ is\ a\ form\ of\ local\ field.$$

VI. Experimental Setup and Simulation

A. *Dataset*

The dataset used for this paper is a combination of printed and handwritten Odia vowel characters which are collected from NIT, Rourkela database for handwritten characters and some downloaded printed vowel characters. The NIT database consists of 47 folders of different Odia characters written by 160 individuals in a specific format with A4 sheet. Each folder contains 320 handwritten samples of dimension 81×81. In this paper we consider all images of vowels contains in the first 11 folders of downloaded database. For printed Odia vowel characters we downloaded 20 different samples of each character type, so a total of 3740 samples with 340 samples per each class are collected. For consonants we consider 100 images from each of 36 consonants folders with a total of 3600 patterns for training and testing in our first implementation and divide these patterns with *k*-fold cross validation method for second part of implementation.

B. *Experimental Setup*

The experimental environment used in this paper is Window8.1 as OS with 16GB memory and Intel $i7$ runs in MATLAB2014a. Two experimental scenarios are presented in this paper. First experiment is to recall all the patterns with different noise percentages from the network. These noises are explicitly introduced with five variations such as 10%, 20%, 30%, 40% and 50% while recall the stored patterns. Two sample images of Odia vowel characters are shown in Fig. 6 with different noise percentages from 10 to 50 percentage of noise .In our second experiment; we partitioned the data with k-fold cross validation method for different training and testing dataset and then recognize test patterns with various noise percentages. To reduce the processing time of the network, image sizes are resize to 30×30.

C. Implementation and Discussion

Before training the network with three mentioned learning rules, the dataset is divided into training and testing and pre-processed. Then we applied histogram equalization, binarization and finally resize them to 30×30 to speed up the training process. For the first implementation we trained the full dataset and tested all patterns for recalling whereas for the second implementation the full dataset is shuffled and then partitioned in training-testing dataset with $k\,(=5) - fold\ cross\ validation$ method. Both the experiments were implemented for three mentioned learning rules with various noise percentages. The training (storage) algorithm is discussed in algorithm-1 for three learning rules to get the weight matrix. The final weight matrix obtained from the *algorithm-1* is used to predict the class labels of test data for each learning rule and the recalling algorithm is given in *algorithm-2*. To predict a test class label hamming distance metric has been considered in this study that finds the minimum hamming distance difference between the test pattern and one of the trained patterns.

Algorithm-1: Hopfield_Training ()

Input:

Input training Dataset *X*

Learning Rule *rule*

Output:

Weight Matrix *W*

1. Shuffle dataset *X*
2. Split X into training and testing dataset (*Full dataset for training and testing in first experiment and k-fold data partition for second experiment)*
3. Get the training dataset *train*
4. Initialize the weight matrix to zero $W_{ij} = 0, \forall\, i, j$
5. Repeat following steps for all the training patterns
6. **if** *rule* is "Hebbian" then
7. perform matrix multiplication of first pattern P_1 with its transpose and obtained the new weight matrix by equation (6). Finally get the weight matrix of all patterns by equation (9).
8. **else if** *rule* is "Pseudo-Inverse" then
9. Obtained the weight matrix by equation (14)
10. **else if** *rule* is "Storkey" then
11. Obtained the weight matrix by equation (16)
12. **end if**
13. **end** step-5
14. Assign the diagonal of the final weight matrix to zero and normalize the weight matrix by dividing with N.

Algorithm-2: Hopfield_Recall/Recognize ()

Input:

Testing dataset *test*

Weight Matrix *W (As per Learning rule)*

Activation Function Y_0

Output:

Output Pattern Y_i

Model Accuracy *In %*

1. Load the weight matrix *W (As per Learning Rule)*
2. Set *correct_pat=0*
3. Set *total_test=0*
4. **for** all test patterns in *test* **do**
5. Assign a test pattern to variable $t = test_i$, *for i=1 to all test patterns*
6. **repeat**
7. equation (2)
8. equation (3)
9. **until** network is stable
10. Assign the output of step-9 to *ypredict*
11. Obtained the minimum *Hamming_distance of ypredict* with all the trained patterns and assign to Y_t
12. Assign Y_{true} = actual label of test pattern *t*
13. **if** $Y_t = Y_{true}$ then
14. *correct_pat=correct_pat + 1*
15. **end if**
16. *total_test=total_test + 1*
17. **end for**
18. *Accuracy=correct_pat / total_test*
19. **return** *Accuracy*

D. Implementation-I

In this paper, the first experiment is performed to demonstrate the recall efficiency of the full dataset (3740 Odia vowel patterns) in Hopfield network with three different learning rules such as, Hebbian learning, pseudo-inverse learning and Storkey learning for various noise percentages and a comparative result is shown in Table II. Similarly for consonants, 100 samples of each class (i.e. 3600 patterns) are considered to store into Hopfield network by using three mentioned learning rules and the recalling efficiency with various noise percentages are shown in Table VI. It has been observed that the recalling efficiency of vowels are more than consonants with different noise percentages as many of these consonants are similar rounded shapes in their upper part.

E. Implementation-II

Second experiment is presented to demonstrate the performance of recognition accuracy by splitting the dataset in train and test data with *k*-fold cross validation method for three different learning rules. In *k*-fold cross validation, the dataset is partitioned into *k* randomly chosen equal (or nearly equally) sized subsets (or folds). One fold is used to validate (testing) and other *k-1* folds are used for training in the network. This process of training and testing is repeated for k times such that each subset (fold) of data is used for testing exactly once. In this paper we portioned Odia vowels dataset into 5-fold (*k=5*) with 748 patterns are in each fold and 720 patterns in each fold for Odia consonants. The *k*-fold cross validation with *k=5* is shown in Fig. 5 and for each learning rules we performed validation for five times named model-1, model-2, model-3, model-4 and model-5 so that each segment of data must be tested once. The recognition accuracy of each model and the average accuracy of each learning rules for Odia vowels are presented in Table III, Table IV and Table V with different noise percentages respectively whereas Table VII, Table VIII and Table IX presented here demonstrates the recognition rate for Odia consonants with Hebbian, Pseudo-inverse and Storkey learning rules respectively. Fig. 7, 8 and 9 are presented here to demonstrate the comparison of different models of k-fold validation method for Odia vowels with Hebbian, Pseudo-inverse and Storkey learning rules respectively whereas Fig. 10 demonstrates the comparison graph of all three learning rules. At the same time Fig. 11, Fig. 12 and Fig. 13 presents the comparison for Odia consonants with three learning rules respectively and Fig. 14 represents the comparison graph of three learning rules for Odia consonants.

TABLE. I. COMPARISON OF DIFFERENT PROPOSED MODELS FOR ODIA CHARACTER RECOGNITION

References	**Method Used**	**Claimed Accuracy(In%)**
Chaudhuri et al.	Feature tree classifier, Run-number based Matching	96.30
Debananda Padhi	GA with ANN	94.00
Pradeepta Sarangi et al	Hopfield Network	95.40
Pradeepta K. Sarangi	ANN	85.30
K. Roy et al.	ANN	97.69
Mamata Nayak et al.	Tesseract OCR	100
Bhagirath Kumar et al.	AMA	92.42- 97.87
Rasmi Ranjan Das et al.	BPNN and SVM	83.33(BPNN) 93.40(SVM)
Manoj K. Mahto et al.	Quadrant Mean method	93.20
Peeta Basa Pati et al.	NN, KNN	82.33(NN) 72.27(kNN)
T.K.Mishra et al.	BPNN with DTC & DWT	92(DTC) 82.70(DWT)
S.D.Meher et al.	BPNN	91.24
U.Pal et al.	Quadratic Classifier	94.60
B. Majhi et al.	ANN with Gradient, Curvature features	99.30(Grad.) 95.66(curv.)
T. K. Bhowmik et al.	HMM	90.50
K.Roy et al.	NN and Quadratic	94.81
N. Tripathy et al.	Normalization and thinning Free automatic scheme	97.74
Kalyan S Dash et al.	Hybrid Features with DLQDF & MQDF	98.50 98.40
Indugu Rushiraj et. al.	Shadow, centroid and distance based features	87.60
S. R. Panda et. al.	Template matching with Unicode mapping	97.00
Kalyan S Dash et.al.	BESAC features with Boolean matching	99.35(ISI) 98.9% (IITBBS) 99.48(ISI)
S. Satapathy et. al.	Deep neural autoencode	97.63

TABLE. II. RECALL EFFICIENCY (IN %) OF HOPFIELD NETWORK WITH THREE LEARNING RULES FOR ODIA VOWELS

Learning Rules	**Recall accuracy (In %) with**					
	Different noise percentages					
	0%	***10%***	***20%***	***30%***	***40%***	***50%***
Hebbian	100	100	100	100	98.56	94.22
Pseudoinverse	100	100	100	100	99.50	97.50
Storkey	100	100	100	100	100	99.12

TABLE. III. RECOGNITION ACCURACY (IN %) FOR ODIA VOWELS IN HOPFIELD NETWORK WITH HEBBIAN LEARNING RULE

Model	**Recognition accuracy (In %) with**					
	Different noise percentages					
	0%	***10%***	***20%***	***30%***	***40%***	***50%***
Model-1	91.20	91.20	90.75	88.56	88.00	86.86
Model-2	93.75	93.50	93.02	92.77	91.90	90.99
Model-3	96.66	96.66	96.20	95.83	94.77	94.20
Model-4	95.82	95.80	95.00	94.87	93.99	92.56
Model-5	91.60	91.60	91.33	90.88	90.00	89.66
Average	**93.80**	**93.75**	**93.26**	**92.58**	**91.73**	**90.85**

TABLE. IV. RECOGNITION ACCURACY (IN %) FOR ODIA VOWELS IN HOPFIELD NETWORK WITH PSEUDO-INVERSE LEARNING RULE

Model	**Recognition accuracy (In %) with**					
	Different noise percentages					
	0%	***10%***	***20%***	***30%***	***40%***	***50%***
Model-1	92.20	91.90	91.75	90.56	89.00	88.06
Model-2	94.75	94.50	94.02	93.67	92.83	92.00
Model-3	97.87	97.66	97.20	96.33	95.87	95.20
Model-4	95.89	95.89	95.20	95.07	94.90	93.63
Model-5	92.66	92.33	92.33	92.00	91.45	91.06
Average	**94.67**	**94.45**	**94.10**	**93.52**	**92.81**	**91.99**

TABLE. V. RECOGNITION ACCURACY (IN %) FOR ODIA VOWELS IN HOPFIELD NETWORK WITH STORKEY LEARNING RULE

Model	**Recognition accuracy (In %) with**					
	Different noise percentages					
	0%	***10%***	***20%***	***30%***	***40%***	***50%***
Model-1	92.50	92.50	92.25	91.63	91.00	89.66
Model-2	95.75	95.75	95.02	94.67	94.03	93.22
Model-3	98.50	98.50	97.90	97.33	96.77	96.00
Model-4	96.22	96.22	95.92	95.35	94.89	94.03
Model-5	92.99	92.99	92.63	92.23	92.00	91.86
Average	**95.20**	**95.20**	**94.74**	**94.24**	**93.73**	**92.95**

TABLE. VI. RECALL EFFICIENCY (IN %) OF HOPFIELD NETWORK WITH THREE LEARNING RULES FOR ODIA CONSONANTS

Learning Rules	**Recall accuracy (In %) with**					
	Different noise percentages					
	0%	***10%***	***20%***	***30%***	***40%***	***50%***
Hebbian	100	100	100	97.33	95.56	92.22
Pseudoinverse	100	100	100	98.00	96.50	94.50
Storkey	100	100	100	99.00	98.02	96.22

TABLE. VII. RECOGNITION ACCURACY (IN %) FOR ODIA CONSONANTS IN HOPFIELD NETWORK WITH HEBBIAN LEARNING RULE

Model	Recognition accuracy (In %) with *Different noise percentages*					
	0%	*10%*	*20%*	*30%*	*40%*	*50%*
Model-1	90.20	90.20	89.75	88.66	87.00	85.86
Model-2	92.85	92.50	92.02	90.77	89.20	87.99
Model-3	96.22	95.89	95.20	94.83	94.00	93.20
Model-4	94.82	94.10	93.60	92.87	91.09	90.06
Model-5	92.60	91.60	91.33	90.22	89.30	88.66
Average	**93.35**	**92.85**	**92.38**	**91.47**	**90.11**	**89.15**

TABLE. VIII. RECOGNITION ACCURACY (IN %) FOR ODIA CONSONANTS IN HOPFIELD NETWORK WITH PSEUDO-INVERSE LEARNING RULE

Model	Recognition accuracy (In %) with *Different noise percentages*					
	0%	*10%*	*20%*	*30%*	*40%*	*50%*
Model-1	91.00	90.90	90.01	88.77	88.00	86.10
Model-2	93.22	92.89	92.00	91.02	89.89	88.66
Model-3	97.00	96.66	95.83	94.90	94.00	93.83
Model-4	95.89	95.66	95.00	94.33	93.87	92.02
Model-5	93.00	92.88	91.66	91.00	90.10	89.20
Average	**94.02**	**93.79**	**92.90**	**92.00**	**91.17**	**98.96**

TABLE. IX. RECOGNITION ACCURACY (IN %) FOR ODIA CONSONANTS IN HOPFIELD NETWORK WITH STORKEY LEARNING RULE

Model	Recognition accuracy (In %) with *Different noise percentages*					
	0%	*10%*	*20%*	*30%*	*40%*	*50%*
Model-1	92.06	91.02	90.63	89.20	88.83	87.50
Model-2	93.66	93.66	92.83	91.02	90.87	90.00
Model-3	98.00	98.00	97.02	96.33	95.20	94.00
Model-4	96.00	95.83	95.00	94.66	93.88	92.33
Model-5	94.00	93.22	92.83	91.33	90.67	89.83
Average	**94.74**	**94.34**	**93.66**	**92.50**	**91.89**	**90.73**

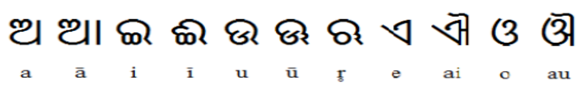


Fig. 1. Odia Vowels with English Transliteration.

Fig. 2. Odia Consonants with English Transliteration.

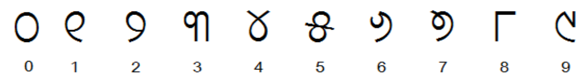


Fig. 3. Odia Numerals with English Numbers.

Fig. 4. Few Similar Odia Characters.

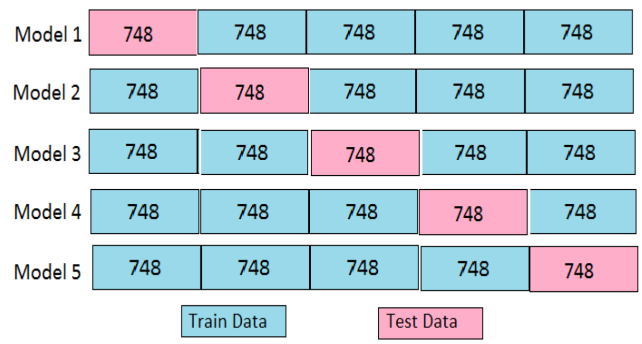


Fig. 5. K-Fold Cross Validation Data Partition for Training and Testing Data with k=5.

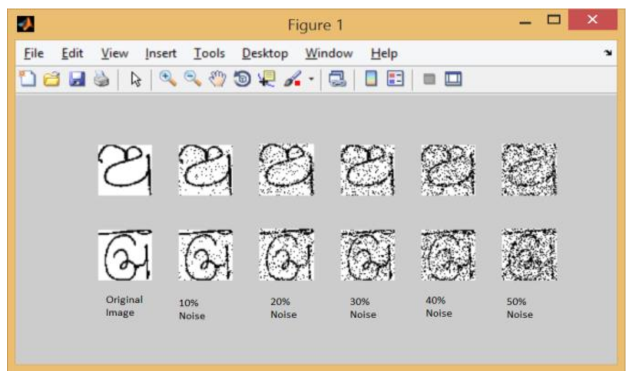


Fig. 6. Sample Images of Two Odia Characters with 10% to 50% Noise.

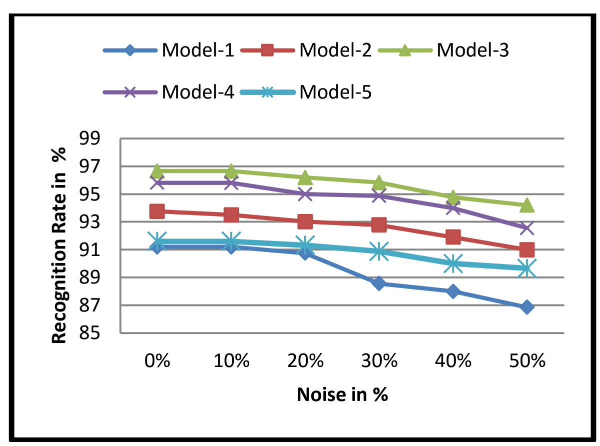


Fig. 7. Recognition Rate of Odia Vowels by K-Fold Cross Validation Model with Hebbian Learning.

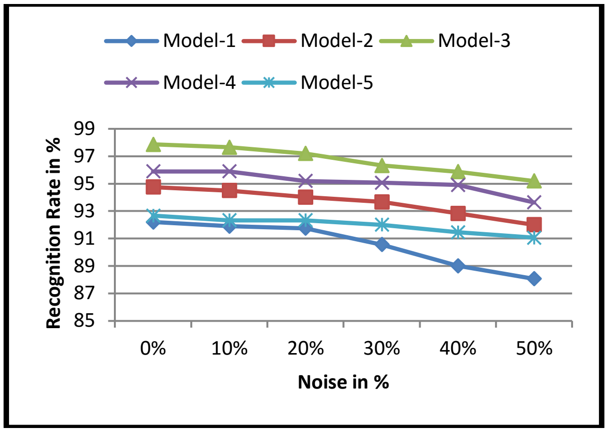


Fig. 8. Recognition Rate of Odia Vowels by K-Fold Cross Validation Model with Pseudo-Inverse Learning.

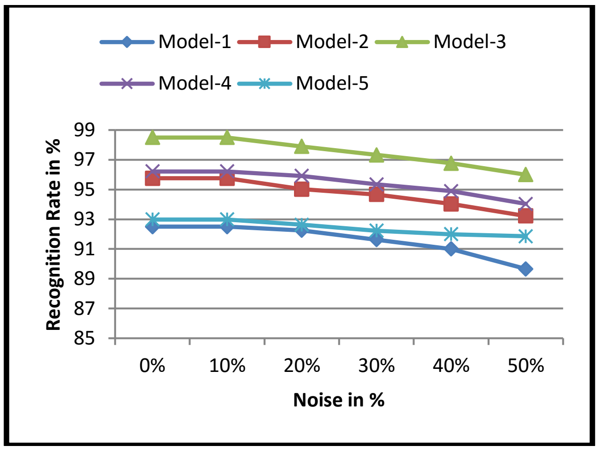


Fig. 9. Recognition Rate of Odia Vowels by K-Fold Cross Validation Model with Storkey Learning.

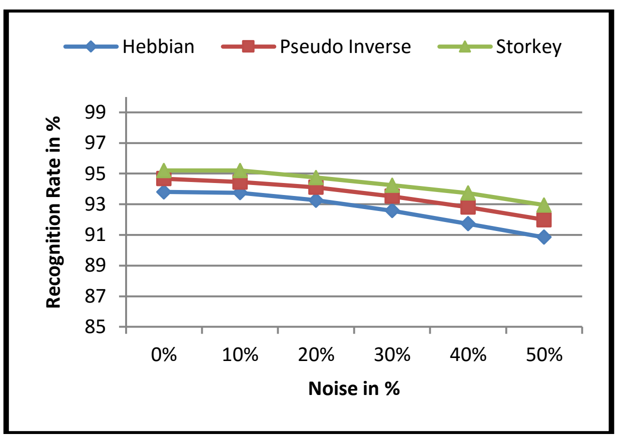


Fig. 10. Comparison Line Graph of Three Learning Rules with Hopfield.

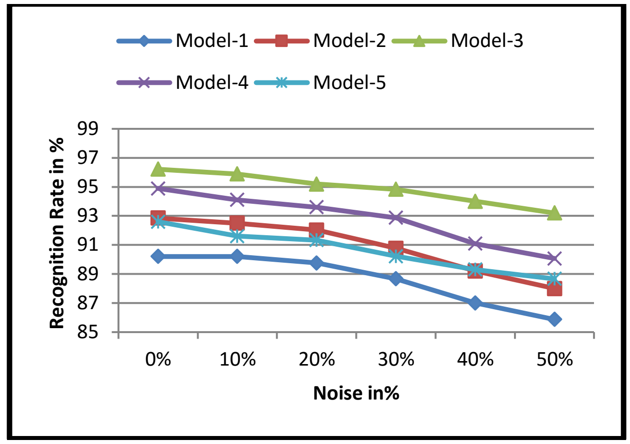


Fig. 11. Recognition Rate (Odia Consonants) of 5-Fold Validation Model for Hebbian Learning Rule.

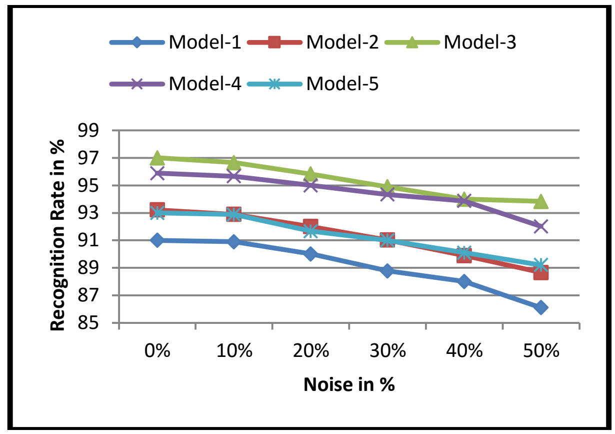


Fig. 12. Recognition Rate (Odia Consonants) of 5-Fold Validation Model for Pseudo-Inverse Learning Rule.

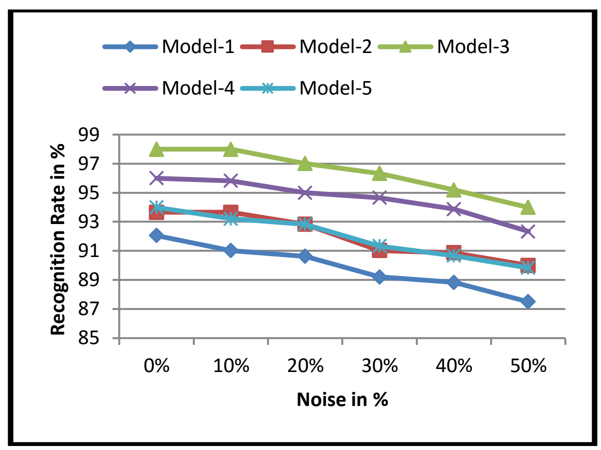


Fig. 13. Recognition Rate (Odia Consonants) of 5-Fold Validation Model for Storkey Learning Rule.

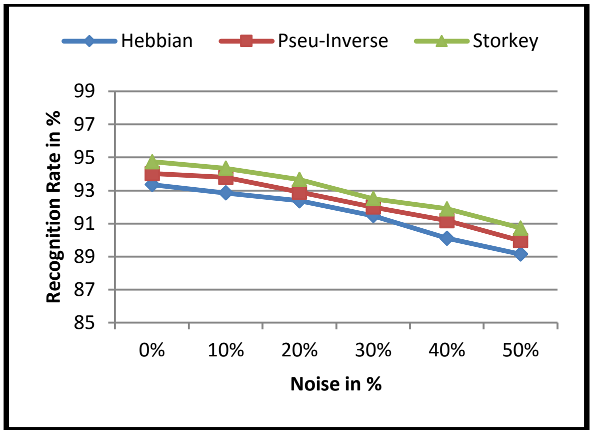


Fig. 14. Comparison Line Graph of Three Learning Rules in Hopfield Network for Odia Consonants.

VII. Conclusion

The present study demonstrated the comparison of three learning rules such as Hebbian, Pseudo-inverse and Storkey rule in Hopfield neural network for recalling and recognizing Odia handwritten and printed basic characters (vowel & consonants) in image form. Our study mainly focused on the recalling performance (stored patterns) and recognizing efficiency (new patterns) for distorted Odia characters. It is hard to compare with other models as very little work has been done so far and no such standardized test set available for Odia character recognition, hence the results we obtained are compared with claimed outputs of different methods in literature survey. The simulation results shows a better accuracy of 95.20% with Storkey learning rule than Hebbian (93.80%) and Pseudo-inverse (94.67%) learning rules without noise for Odia vowels and in case of Odia consonants Storkey rule (94.74%) also out performed than other rules. In this simulation work for recognising a pattern, we used a machine learning validation method named *k*-fold cross validation to partitioning the dataset in five sub parts and in every iteration one sub part is for testing and other four sub parts are used for training by which each pattern must be tested with various noise percentages. It has been seen that Storkey learning rule provides better accuracy in terms of recognising new test sets with noise as compared to other two learning rules.

References

[1] S. Rajasekaran and B. Deekshatulu, “Recognition of printed telugu characters”, Computer Graphics and Image Processing, vol. 6, no. 4, pp. 335–360, 1977.

[2] G. Siromoney, R. Chandrasekaran, and M. Chandrasekaran, “Computer recognition of printed tamil characters”, Pattern Recognition, vol. 10, no. 4, pp. 243–247, 1978.

[3] M. Maloo and K. Kale, “Support vector machine based gujarati numeral recognition”, International Journal on Computer Science and Engineering, vol. 3, no. 7, pp. 2595–2600, 2011.

[4] B. Dhandra, G. Mukarambi, and M. Hangarge, “Zone based features for handwritten and printed mixed kannada digits recognition”, in IJCA Proceedings on International Conference on VLSI, Communications and Instrumentation (ICVCI), no. 7, pp. 5–8, 2011.

[5] D. V. Sharma, G. S. Lehal, and P. Kathuria, “Digit extraction and recognition from machine printed Gurmukhi documents”, in Proceedings of the International Workshop on Multilingual OCR. ACM, no. 12, pp. 1–8, 2009.

[6] B. Chaudhuri, U. Pal, and M. Mitra, “Automatic recognition of printed oriya script”, Sadhana, vol. 27, no. 1, pp. 23–34, 2002.

[7] U. Pal and B. Chaudhuri, “Automatic recognition of unconstrained off-line Bangla handwritten numerals”, in Advances in Multimodal Interfaces—ICMI 2000. Springer, pp. 371–378, 2000.

[8] S. Mohanti, “ Pattern recognition in alphabets of Oriya language using Kohonen neural network”, Int. J. Pattern Recogn. Artif. Intell. Pp.1007–1015, 12 (1998).

[9] K. Roy, T. Pal, U. Pal, and F. Kimura, “Oriya handwritten numeral recognition system” , in Eighth International Conference on Document Analysis and Recognition. IEEE, pp. 770–774, 2005.

[10] T. K. Bhowmik, S. K. Parui, U. Bhattacharya, and B. Shaw, “An hmm based recognition scheme for handwritten oriya numerals”, in 9th International Conference on Information Technology. IEEE, pp. 105–110, 2006.

[11] P. K. Sarangi, A. K. Sahoo, and P. Ahmed, “Recognition of isolated handwritten oriya numerals using hopfield neural network”, International Journal of Computer Applications, vol. 40, no. 8, pp. 36–42, 2012.

[12] J. Panda, M. Panda, A. Samal, and N. Das, “Odia handwritten digit recognition using single layer perceptron”, International Journal Of Electronics And Communication Engineering & Technology, vol. 5, no. 4, pp. 80–88, 2014.

[13] K. S. Dash, N. Puhan, and G. Panda, “A hybrid feature and discriminant classifier for high accuracy handwritten odia numeral recognition”, in IEEE Region 10 Symposium, IEEE, pp. 531–535, 2014.

[14] S. Meher and D. Basa, “An intelligent scanner with handwritten odia character recognition capability” , in Fifth International Conference on Sensing Technology (ICST). IEEE, pp. 53–59, 2011.

[15] B. Kumar, N. Kumar, C. Palai, P. K. Jena, and S. Chattopadhyay, “Optical character recognition using ant miner algorithm: a case study on Oriya character recognition”, International Journal of Computer Applications, vol. 61, no. 3, pp. 17–22, 2013.

[16] T. K. Mishra, B. Majhi, P. K. Sa, and S. Panda, “Model based odia numeral recognition using fuzzy aggregated features”, Frontiers of Computer Science, vol. 8, no. 6, pp. 916–922, 2014.

[17] P. Pujari and B. Majhi, “A comparative study of classifiers on recognition of offline handwritten odia numerals” , in International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO). IEEE, pp. 1–5, 2015.

[18] Chen, L., Fan, J. and Chen, Y. , “A High Speed Modified Hopfield Neural Network and A Design of Character Recognition System”, IEEE Chung-Yung Christian University, CH3031-2/91/0000-0308, pp.308-314, 1991.

[19] M.B. Sukhaswami, P. Seetharamulu and Arun K. Pujari “Recognition of Telugu Characters using Neural Networks”, International Journal of Neural Systems VOL. 06, NO. 03.

[20] Prateek Malik ; Rinku Dixit “Handwritten Character Recognition Using Wavelet Transform and Hopfield Network” , International Conference on Machine Intelligence and Research Advancement, IEEE, 2013.

[21] Hopfield, John J. "Neural networks and physical systems with emergent collective computational abilities." Proceedings of the national academy of sciences, pp.2554-2558, 79.8 (1982).

[22] Hebb D “The Organization of Behavior. A Neuropsychological Theory”. Wiley, New York, NY, 1949.

[23] Personnaz, L., I. Guyon, and G. Dreyfus, “Collective Computational Properties of Neural Networks: New Learning Mechanisms”, Physical Review A, 1986. 34(5): pp. 4217-4228.

[24] Storkey, Amos. "Increasing the capacity of a Hopfield network without sacrificing functionality", Artificial Neural Networks – ICANN'97, pp.451-456, 1997.

[25] Kalyan S Dash, N. B. Puhan, Ganapati Panda, “A hybrid feature and discriminant classifier for high accuracy handwritten Odia numeral recognition”, 2014 IEEE Region 10 Symposium, pp. 531-535, 2014.

[26] Kalyan S Dash, N. B. Puhan, G. Panda, “Odia character recognition: a directional review”, Artificial Intelligence Review, Springer, Vol. 48(4), pp. 473-497, 2016.

[27] Indugu Rushiraj, Souvik Kundu and Baidyanath Ray, "Handwritten Character Recognition of Odia Script", International conference on Signal Processing, Communication, Power and Embedded System (SCOPES), pp. 764-767, 2016.

[28] Smruti Rekha Panda and Jogeswar Tripathy, " Odia Offline Typewritten Character Recognition using Template Matchingwith Unicode Mapping", International Symposiwn on Advanced Computing and Communication (ISACC), 2015.

[29] Kalyan S Dash , N.B. Puhan and Ganapati Panda, " BESAC: Binary External Symmetry Axis Constellation for unconstrained handwritten character recognition", Pattern Recognition Letters, Elsevier, Vol. 83(3), pp. 413-422, 2016.

[30] Subhashree Satapathy et. al., "Printed Odia Numerals Recognition using Stacked Autoencoder", 8th International Congress of Information and Communication Technology, ICICT 2019, Procedia, Computer Science, pp.790-797, 2019.

Critical Analysis of Brain Magnetic Resonance Images Tumor Detection and Classification Techniques

Zahid Ullah[1], Prof. Su-Hyun Lee*[2], Donghyeok An[3]
Department of Computer Engineering
Changwon National, Changwon
South Korea

***Abstract*—The image segmentation, tumor detection and extraction of tumor area from brain MR images are the main concern but time-consuming and tedious task performed by clinical experts or radiologist, while the accuracy relies on their experiences only. So, to overcome these limitations, the usage of computer-aided design (CAD) technology has become very important. Magnetic resonance imaging (MRI) and Computed Tomography (CT) are the two major imaging modalities that are used for brain tumor detection. In this paper, we have carried out a critical review of different image processing techniques of brain MR images and critically evaluate these different image processing techniques in tumor detection from brain MR images to identify the gaps and limitations of those techniques. Therefore, to obtain precise and better results, the gaps can be filled and limitations of various techniques can be improved. We have observed that most of the researchers have employed these stages such as Pre-processing, Feature extraction, Feature reduction, and Classification of MR images to find benign and malignant images. We have made an effort in this area to open new dimensions for the readers to explore the concerned field of research.**

Keywords—Magnetic Resonance Imaging (MRI); Computed Tomography (CT); MRI classification; tumor detection; digital image processing

I. INTRODUCTION

Due to abnormal cell development, the brain tumor begins and grows in an uncontrolled way. The brain cells of a human brain can specifically ruin by the tumor. Different brain tumors are illustrated in Fig. 1. The most dangerous form of tumors is malignant tumors. Every year fourteen thousand deaths caused due to malignant tumors. The tumor stages are divided into different grades due to severity levels such as grade 1, grade 2, grade 3, and grade 4. The grade 1 level is the least dangerous tumor and this type of tumor grows slowly and gradually. For this type of tumor grade, treatment via surgery might be successful. Ganglioglioma, gangliocytoma, and pilocytic astrocytoma are the different cases of grade 1 brain tumor. In the second grade, grade 2 tumor also grows slowly [24], and this type of tumor looks irregular using the microscopic instrument. The third one is grade 3 tumor which is also malignant and there is no significant difference between grade 3 and grade 2. The maximum malignant tumor is grade 4 and the example of grade 4 tumor is Glioblastoma Multiforme [25].

Different image processing techniques have been used for tumor detection. Segmentation of images is one of them. The image segmentation aim is to segment an image into equal parts and find the region of interest (ROI) [26-27].

Nowadays, various imaging modalities are available such as Magnetic Resonance Imaging (MRI), X-Ray, Positron Emission Tomography (PET), Ultrasound, and Computed Tomography (CT) scan. In the given imaging modalities, magnetic resonance imaging generates good quality images and these images are extremely helpful for clinical diagnosis and biomedical research. MRI is particularly useful in examining the soft tissues in the human body.

Computed Tomography (CT) scan and Magnetic Resonance Imaging (MRI) are the two most useful imaging modalities to examine brain life structure. However, Magnetic resonance imaging is currently a widely used method as it produced high-quality medical images among all the imaging modalities. MRI is a non-invasiveness technique and it discriminates the soft tissue [28]. Magnetic resonance imaging (MRI) gives accurate information on the internal human body. MRI demonstrates the physiological or pathological variations of human tissues. MRI is mostly utilized when treating prostate cancers, brain, foot, and ankle. MRI can be used for the identification of different diseases, such as Alzheimer's, Parkinson, Stroke and Brain tumors diseases. The performance of MRI is far better than a CT scan. The MRI works effectively for soft tissues in the body, such as the brain, while the CT scan is suitable for hard body tissues such as bones.

Furthermore, image segmentation is a necessary task which is nowadays carried out manually by most physicians as they obtained high accuracy, but it requires a high amount of time which is the main drawback. However, on the other side, automatic segmentation techniques are yet not reliable. Therefore, to overcome these issues, the semi-automatic segmentation techniques are currently being used for clinical applications which are the best way to solve the limitations of automatic segmentation simultaneously. The user initialization is required in several semi-automatic methods to assure the accuracy rate.

*Corresponding Author (sleepl@changwon.ac.kr)

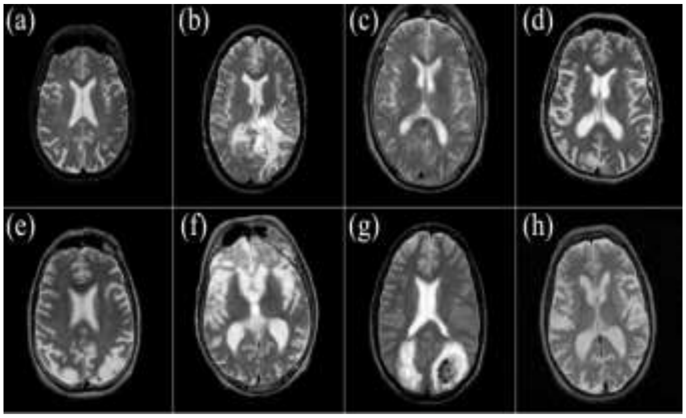


Fig. 1. Different Brain Diseases (a) Normal Brain; (b) Glioma; (c) Meningioma; (d) Alzheimer's Disease; (e) Alzheimer's Disease with Visual Agnosia; (f) Pick's Disease; (g) Sarcoma; (h) Huntington's Disease.

The appropriate segmentation of the MR image tumor part improves the efficiency of the techniques. The researcher must have adequate knowledge about the image processing techniques to process different kinds of medical images. Therefore, for this purpose, we have reviewed and critically analyzed different tumor detection techniques in MR images used by various researchers to identify the gaps and limitations.

The structure of the rest of the review paper is as follows: Section 2 presents the Literature Review, Section 3 presents computer vision and image processing, Section 4 presents Critical Analysis of different existed methods, Section 5, presents Discussion and Section 6 presents Conclusion and Future Work.

II. LITERATURE REVIEW

In this section the existing techniques of brain tumor segmentation and classification have been discussed. The conventional approach to identify the brain abnormality is manual, which is very prone to misclassify and is time consuming. Therefore, researchers have proposed numerous automated techniques to overcome this issue.

The processing and analyzing of brain tumor MR images are the most challenging task. Different researchers have proposed different techniques to classify the brain tumor in MR images such as support vector (SVM), fuzzy clustering mean (FCM), artificial neural network (ANN), expectation-maximization (EM) and knowledge-based technique algorithm. These mentioned algorithms are the most popular algorithms which are employed for region-based segmentation to extract the required data from MR images. We have critically reviewed some of the proposed techniques to pave the way for further research.

El-Dahsan et. al., [1] proposed a hybrid technique to classify the magnetic resonance images. The proposed algorithm comprised of three stages, such as feature extraction, feature reduction, and classification. In the first stage, Discrete Wavelet Transform (DWT) is employed, then Principal Component Analysis (PCA) is used for feature reduction in the second stage and lastly, in the third stage of the methodology, Feed Forward Back-Propagation Artificial Neural Network (FP-ANN) and k-Nearest Neighbor (k-NN) algorithm are employed for the classification process. A total of 70 (60 abnormal and 10 normal) axial T2-W images were used and achieved 97% and 98.6% accuracy on ANN and k-NN algorithm, respectively.

Nazir et al., [2] proposed an automatic approach for brain MRI classification. Three stages have been used in the methodology, such as pre-processing, feature extraction and classification. In the pre-processing step, the median filter is applied for noise removal from the MR image. The color moments are extracted as mean features from the MR images in the feature extraction stage. The feed-forward artificial neural network has been used to classify these extracted features as normal or abnormal. In the proposed methodology, they have used 70 (45 abnormal and 25 normal) T2-W MR images and obtained a 91.8% accuracy rate.

Lavanyadevi et al. [3], used a neural network to classify the brain tumor phase such as malignant, benign or normal. Gray Level Co-occurrence matrix is used for feature extraction. Principal component analysis has been used for dimensionality reduction while the classification stage of the MR brain image has been done by using a probabilistic neural network and achieved some satisfactory results.

Amin et al., [4] classify the normal and abnormal brain MRI. In the pre-processing stage, the non-cerebral tissues are eliminated from MR images using morphological operation; afterward, the Gaussian filter is applied for noise removal. Then for the feature set, they have chosen texture, shape, and intensity. Finally, the support vector machine is employed to classify the normality and abnormality.

Kumar et. al., [5] presented a hybrid approach for brain tumor detection. For feature extraction and reduction, discrete wavelet transform (DWT) and a genetic algorithm have been used respectively. A support vector machine (SVM) has been employed for the classification of a brain tumor as benign or malignant. A data set of 25 (20 normal and 5 abnormal) T2-W brain MR images were obtained for the SICAS medical image repository. Different parameters have been used for analyzing these images such as root mean square error (RMS), smoothness, and entropy and achieved some satisfactory results.

Iscan et al., [6] mainly focused on tumor detection in MR brain images by finding asymmetry in the right or left brain hemisphere. They have used an incremental supervised neural network for the segmentation of MR images. For image brightness, continuous wavelet transform has been applied. While Zernike is applied for vector representation, this method obtained 100% accuracy on both normal and abnormal MR image segmentation.

Zhang and Wu [7] used the wavelet transform for feature extraction from MR images. Principal component analysis has been used for feature reduction. Then these reduced features were sent to the kernel support vector machine. They have selected seven different diseases. 20 normal and 140 abnormal images were collected from Harvard medical school website and achieved some good results.

Fayaz et al., [8] used a three steps algorithm for brain MRI classification such as pre-processing, feature extraction, and classification. The median filter is applied to a grayscale MR image for noise removal and then converted this grayscale image to an RGB format in the first stage. The three channels were extracted i.e., red, green and blue from every single channel of RGB in the feature extraction stage, then classified these extracted features as normal or abnormal in the third stage. They have used 100 T2-weighted images (30 abnormal and 70 normal) and obtained some good results.

Zarendi et. al., [9], proposed a method for tumor detection in the brain MR image by using the type-II fuzzy technique. The fuzzy logic has been used in some specific areas to solve the uncertainty about the data classification. Demirhan and Guler [10], have combined stationary wavelet transformation with a self-organization map to segment the MR images. The standard images have been considered from the internet brain segmentation repository (IBSR) and the MRI experts have performed the results of segmentation.

Ibrahim et al., [11], used principal component analysis (PCA) to reduce the dimension of the data. While for the detection of a brain tumor in MR images, the supervised feed-forward back-propagation neural network has been used. They have used three layers of ANN such as, i) the input layer which contains 64 neurons ii) hidden layers which comprised of 10 neurons and iii) the output layer which consists of 64 neurons. They have classified the images in four different classes such as normal tissues, Edema, cancerous tissue, and not a classified class. This proposed methodology achieved a 96.33% accuracy result.

Jafarpour et. al. [12], using the gray level co-occurrence matrix in MR images for feature extraction. The two most essential algorithms were combined in a feature reduction stage to select the best features. Then those features were presented to the artificial neural network and k-nearest neighbor and obtained a 100% accuracy rate on both normal and abnormal images.

Saha et. al., [15] presented a score based bounding box technique for approximate segmentation from MR brain image. In 2D slices, the MR images were segmented and the region based global change has also been observed. The image tumor part is measured as a change in MRI. The Fast bounding Box technique is used to find the asymmetric region in MR image slices. 92% accuracy for tumor detection and 89% accuracy for Edema detection been observed. Węgliński and Fabijańska [17], have used region growing algorithm and extracted the complete abnormal tissue. For separation of the normal and abnormal area, the seed point is employed. The median filter is used for noise removal as it preserves the edges and does not affect the image quality. The methodology produced some satisfactory results. The pre-processing and post-processing stages increase the execution time.

Shanthi and Sasikumar [21] have employed fuzzy logic and neural network algorithm for MR image segmentation. In the image segmentation, they have mainly focused on white matter, gray matter, and cerebrospinal fluid. Fuzzy logic has been used for clustering, and the output of the fuzzy system is presented to the artificial neural network as an input. For the testing purpose, they have considered T1- weighted images.

Shree and Kumar [22] used morphological filtering for noise removal to enhance the MR brain image and used gray-level co-occurrence matrix (GLCM) for feature extraction and discrete wavelet transform (DWT) has been used for feature reduction to reduce the complexity. Lastly, for the detection of tumor location in the MR brain image probabilistic neural network (PNN) classifier has been employed and achieved 100% and 95% accuracy rate for training and testing dataset respectively.

Bahadure et al., [23], an automated methodology has been proposed for brain MR image tumor segmentation. The methodology employed Berkeley wavelet transform and SVM for feature extraction and classification, respectively. They obtained 96.51%, 94.2% and 97.72%, accuracy, specificity, and sensitivity, respectively.

III. Computer vision and Image Processing

The ultimate goal of computer vision is to employ computers to imitate human vision. The image analysis field lies between image processing and computer vision [29]. The high, mid and low-level processes are the general computerized processes. The high-level processing is connected with vision. Whereas, the segmentation and classification lie in mid-level processing. The low-level processing does noise reduction and image contrast enhancement.

The MRI image contained a high amount of information like for instance, the current MRI system can generate images that are equals to 65,535 gray-level [30]. Extracting all the information from this MR image by using human vision is not possible. Because the human eye is unable to differentiate between thousands of gray-levels. Therefore, the use of a computer is the best option to understand and evaluate the high-resolution images in depth.

The classification of brain MRI images follows different stages, such as pre-processing, feature extraction, feature reduction, and classification as shown in Fig. 2. These stages are thoroughly explained in sections below.

A. Pre-Processing

Different variety of pre-processing methods are applicable for different circumstances such as adaptive, linear, non-linear, multi-scale or pixel-based [31]. The noise appears in the MR image because of the variation of the magnetic field in the coil [32]. The noise, partial volume effect, and shading artifact are generally associated with MR images. There is an essential connection between Signal to Noise Ratio (SNR) and Contrast to Noise Ratio (CNR) in MRI application [33]. For diagnostic purposes, high spatial resolution and high contrast are important. Most of the algorithms are noise sensitive; therefore, a high SNR is mandatory in image processing applications. In this regard, noise filtering techniques are required to apply on MR images to remove the noise and preserve the edges of the image. There are various approaches that have been used to improve the SNR and CNR via wavelet filters, adaptive filters, and anisotropic diffusion filters. Out of these techniques, the anisotropic diffusion filtering method illustrates some good results due to its computational speed and its simplicity.

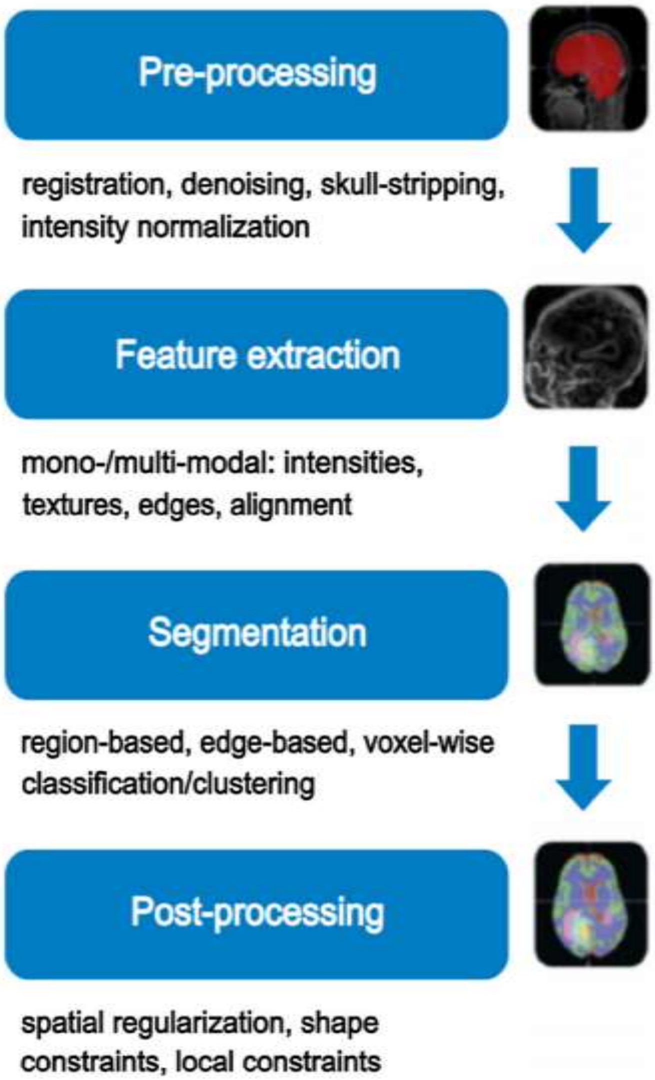


Fig. 2. Schematic Diagram of Brain MRI Segmentation.

The medical images are affected by different types of noises. The MRI images are spoiled by salt and pepper noise, Gaussian noise, speckle noise, etc. It has been proved that for noise removal from MR image, a median filter is the best choice as a linear filter [34]. The focus on the MR image pre-processing stage is essential before it is fed to the classifier; otherwise, we will get the wrong results [35].

Image enhancement is another step required in the pre-processing stage. The different method has been used in [36] for MR image enhancement such as contrast stretch normalization, histogram stretching, histogram normalization, Gaussian kernel normalization, intensity scaling, and histogram equalization. By comparing the results of these techniques, the histogram normalization technique gives the best performance.

The background of an image does not provide any information, and it also maximizes the computation time [37]. So to decrease memory usage and enhance the processing speed, the skull, scalp, eyes, and background needs to be eliminated. The Brain Surface Extractor in [37],[38] has been used for MRI skull stripping.

B. Feature Extraction

Feature extraction is the process of converting an image into its group of features. Mostly these techniques are used for feature extraction such as Gabor features, Bag-of-words, texture features, co-occurrence matrix, wavelet transform based features, etc. In [39], discrete wavelet transform and bag-of-words techniques are used to extract the features from brain MRI. Recognition of essential features leads us to design an efficient system. With feature extraction, the time, data and memory get decreased [40]. The feature extraction stage is extremely important as the outcome is calculated based on these extracted features data [41]. Efficient feature selection makes accurate feature set by eliminating extra features and also overcome dimensionality issues [42]. Essential features are obtained from brain MR images using Fourier transform, independent component analysis and wavelet transform [43].

Nowadays, most researchers used a combination of spectroscopic or MRS features to discriminate against the type of brain tumor. The MRS features provide precise results as mention in [44]. Some techniques depend on high dimensional features; therefore, high memory storage is required [45]. Fig. 2 illustrates the schematic brain MRI segmentation.

Techniques such as Independent component analysis, linear discriminant analysis, and principal component analysis are used for feature reduction. The combination of feature extraction algorithm with feature reduction algorithm takes us to the perfect system that utilizes limited features which can be extracted with low computational cost [42].

C. Dimensionality Reduction

In this stage, the dominant features are selected from the extracted features stage using different techniques such as principal component analysis, independent component analysis, and genetic algorithm. Feature reduction aim's to reduce the size of the dataset and select the essential features to reduce computational time and complexity. Considering all the features of an image badly affect the system accuracy. Hence, to design a robust system to classify the brain MR image accurately is to consciously select the features using appropriate techniques that decrease unimportant information. Because the efficient technique employs a low number of features. However, extensive reduction of features may reduce the system accuracy [47]. Model efficiency is improved by selecting features. Hence avoiding this stage leads us to poor classifier performance [48].

It is also notable from the literature review that the most popular technique for feature reduction is principal component analysis. This technique transforms the input feature space into a lower-dimensional feature space using the eigen vector correlation matrix [49]. In [50], PCA has been used to reduce the extracted feature from 65536 to 1024 features and obtained a high accuracy rate.

Furthermore, the classification accuracy is enhanced using the dimensionality reduction stage. The feature extraction and feature reduction play an essential part in MRI image classification.

D. Classification using Supervised Methods

In this stage, supervised techniques such as ANN, k-NN, SVM, etc. and unsupervised techniques such as Self-organizing maps and k-means techniques are finally used to classify the MRI image into normal or abnormal. The accurately labeled data is used in supervised techniques that are acquired in the training stage so that to determine the class for unlabelled data of the testing phase [55]. The performance of the supervised classifier is better than an unsupervised classifier in terms of the accuracy rate. All these mentioned algorithms require these stages we discussed above to classify an MRI image. However, the Convolutional neural network does not require to follow these mentioned stages as CNN learns its features automatically from the image by using trainable convolutional filters [51]. We have discussed a few classifiers used in different proposed brain MRI classification methodologies:

E. Artificial Neural Network

In [46] employed ANN for brain tumor segmentation. The proposed algorithm utilized GLCM+GA+NN for feature extraction, feature reduction and classification, respectively. A neural network classifier is used in [52] to classify the MR image into normal or abnormal. The features were extracted using cubic order differentiations whereas, rule technique was employed for feature selection. Two classifiers were used for brain MRI classification namely feed-forward backpropagation network and k-NN in [53]. They have extracted the features using discrete wavelet transform and the features were reduced from 1024 to 7 by PCA. Further, in reference, [54] feature extraction task is done by using PCA and MRI image classification is carried out by PNN. A hybrid machine learning approach is used in [56] for brain tumor detection. The proposed methodology comprised of pre-processing, feature extraction, feature reduction and classification. The median filter is used for noise removal, DWT and PCA are used feature extraction and feature reduction respectively. Finally, the feed-forward BPNN is used to find the normality and abnormality of the MRI.

F. K-Nearest Neighbor

This algorithm is a supervised learning algorithm, and it compares the new unlabelled problem instances directly with the labeled samples in the training set. In [57] k-NN algorithm is employed for brain MRI image classification. In the proposed methodology, the median filter is used in the pre-processing stage for noise removal. The color moments such as mean, variance, and skewness are used for feature extraction. The dataset comprised of 100 brain MRI images in which 70 were normal and 30 were abnormal and obtained reasonable accuracy.

Furthermore, in [58] k-NN is employed to segment dark and light abnormalities in both low background and medium background gray level values in MRI images. The k-NN algorithm is used for brain MRI classification in [65]. DWT is used for feature extraction and PCA is employed for feature reduction. In their proposed methodology, they have measured and compared seven different statistical techniques such as kurtosis, skewness, and specificity etc.

G. Support Vector Machine

The SVM can represent complex surfaces including radial basis function and polynomials. The largest margin between the two classes is considering the best hyperplane. The data points that are closest to the separating hyperplane are known as support vectors [59], [60]. This algorithm split the image into two different classes and find the best hyperplane to classify the data efficiently. SVM and GA with Gaussian radial basis functions are used in [61] for brain tumor classification and obtained reasonable accuracy. The SVM classifier is used in paper [62] to classify the tumor type and grade. They have also calculated specificity and sensitivity which provide good results. In [63] presents a methodology for brain MRI classification which is comprised of GA, spatial gray level dependence method and SVM, and they obtained a high accuracy rate. The combination of multi-resolution independent component analysis (MICA) and SVM [64] has increased the accuracy rate 2.5 times higher as compare to other ICA based classification.

H. Unsupervised Techniques

Unsupervised techniques do not require labeled data. The unsupervised algorithms automatically determine the number of classes. These types of algorithms can solve complex problems. Some of the unsupervised algorithms used by different researchers for brain MRI classification are discussed as:

I. Self-Organizing Maps

In [66], an automated hybrid self-organizing maps with fuzzy k-means was proposed to identify normal and abnormal MRI image. They have used clustering for the segmentation process. The greedy k-means algorithm in FKM provides faster convergence. This hybrid algorithm exactly identified the dimension of the tumor region. Brain MRI images are classified into normal or abnormal by using neural network SOM and SVM is used in [43] and acquired 94% and 98% accuracy for SOM and SVM respectively. Another unsupervised k-means algorithm was implemented in [67] for image segmentation purposes. Partial stretching enhancement is done in the pre-processing step, while median filtering is used for noise removal in the post-processing step and they obtained good segmentation results.

J. K-Means Clustering

Clustering is an unsupervised method for image classification. There are two different types of clusters i.e. soft clustering and hard clustering. The data point may belong to a cluster or may not in hard clustering and also the clusters do not overlap. Whereas in soft clustering, the clusters can overlap. Therefore, a single data can belong to two or more than two clusters. K-Means and hierarchical clustering are employed for tumorous tissue separation purposes from an MRI image in [68]. The pseudo-color transformation method is applied to the MRI image and then converted this MRI image into an RGB image for feature enhancement. The same methodology is also implemented in [69] and obtained good results. The complete methodology of brain MRI classification has been shown in Fig. 3 and the general brain structure is depicted in Fig. 4.

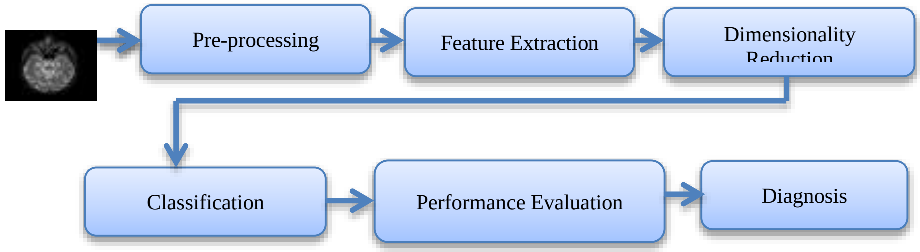


Fig. 3. Methodology of Brain MRI Classification.

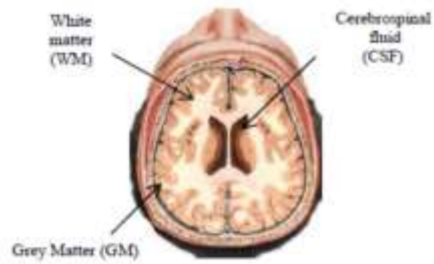


Fig. 4. Brain Image.

IV. Critical Analysis

We have critically reviewed and summarized the literature review part in Table I.

A. Performance Evaluation

The classification accuracy can be evaluated using several ways. We observed from the literature review that confusion matrix, peak signal to noise ratio, Jaccard Index, Accuracy, Sensitivity, Specificity, and Mean Square Error are the most popular statistical techniques to measure the classifier performance.

TP illustrates - Correctly classified positive instances,

TN illustrates - Correctly classified negative instances,

FP illustrates - Incorrectly classified negative cases, and

FN illustrates - Incorrectly classified positive instances.

- Sensitivity

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

- Specificity

$$Specificity = \frac{TN}{TN + FP}$$

- Accuracy

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

B. Datasets

For training and testing purpose various, datasets are being used. Such as, BRATS 2012, 2013, 2015, 2016, 2017, 2018 [76-78], Harvard [78], SPL database [79], PGIMER dataset [79], MRBrain S Challenge dataset [80], ISLES 2015 [81], BrainWeb [82], and so on. Different brain tumor datasets are illustrated in Fig. 5 utilized by researchers to validate their methodologies.

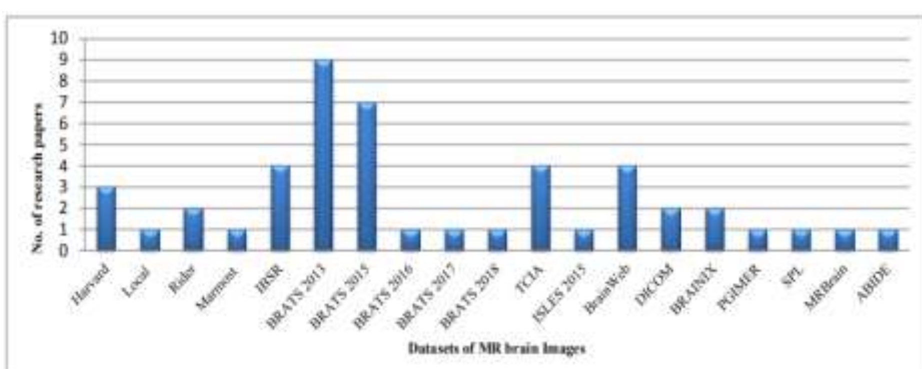


Fig. 5. Depicts Brain MRI Datasets used in Various Publications.

TABLE. I. CRITICAL ANALYSIS OF DIFFERENT PROPOSED MRI METHODOLOGIES

Ref-erences	Focus Area/Feature Used	Algorithm/Technique	Strengths	Limitations	Experimental Results
[1]	-Brain MRI classification -A discrete wavelet transform is used for feature extraction.	-Discrete wavelet transform -Principal Component Analysis (PCA) -Feedforward back propagation artificial neural network (FP-ANN) and k-nearest neighbor(k-NN)	- ANN's physical implementation is simple and straightforward. -The ANN generates accurate results of generalization property. - and can map the distribution of complex classes easily. While k-NN is a suitable choice for smaller data.	-An extremely reduction of features from1024 to 7 may lose some important information.	-ANN obtained 90% accuracy. -k-NN obtained 99% accuracy
[2]	-Color moments -Mean -Variance -Skewness	- Artificial neural network -median filter -arithmetic mean -Standard deviation -Skewness	-The proposed method obtained good results due to the simplicity of the methodology	-The processing time got increased due to pre-processing and post-processing stages.	- The proposed methodology execution time is less than 20s.
[3]	-Brain tumor detection using Gray level co-occurrence matrix	-The gray level co-occurrence matrix -Principal component analysis -probabilistic neural network -K-means clustering	-The probabilistic neural network quickly learns from training data and can adapt its learning in real-time.	-The computation time is high due to many proposed stages.	-The methodology is robust; however, the accuracy has not been mentioned.
[4]	- Texture, shape, and intensity - The gray level co-occurrence matrix	-Morphological operation -Gaussian filter -Support Vector Machine	-The performance of the support vector machine is notable even on large datasets	- In pre-processing two steps have been used such as skull stripping and noise removal which consume extra time	-A total of 100 MR images was used in which 65 were abnormal and 35 were normal. -The obtained accuracy rate is 97.1% - The specificity rate is 98.0% - Sensitivity rate is 91.9%
[5]	-Smoothness, entropy, root mean square error, correlation, and kurtosis are the parameters used in this paper. -A genetic algorithm is used for feature extraction.	-Discrete wavelet transform -Genetic algorithm - Principal component analysis -Support vector machine	-The support vector machine is a powerful classification tool and can be used for both linear and non-linear data. - A genetic algorithm is effective to find the solution to search and optimization problems.	Support vector machine works effectively, however, the accuracy of SVM gets affected on small datasets	-T2-weighted brain MR images dataset has been used which comprised of 25 images. - 20 normal and 5 abnormal. - The accuracy rate is not clearly mentioned
[6]	-Combined the segmentation method and mid-sagittal plane extraction method to detect tumor and asymmetry.	-2D continuous wavelet transform -incremental supervised neural network -Zernike moment by the vector representation -Euclidean distance.	-The artificial neural network implementation is easy and easily map the complex distribution. ANN generates efficient results.	-The Zernike moment's calculation is complex. -The algorithm takes wrong decisions when the small asymmetric differences increase the value of ND2.	-100% accuracy has been obtained for 50 normal brain MR images and 20 abnormal brain MR images.
[7]	-Seven common diseases such as Meningioma, glioma, Pick's disease, sarcoma, Huntington's disease, Alzheimer’s disease, Alzheimer's disease plus visual agnosia has been considered. -Discrete Wavelet Transform is employed for feature extraction	-Discrete Wavelet Transform (DWT). -Principal component analysis (PCA). -Kernel support vector machine (KSVM) with GRB kernel.	-The support vector machine is very effective for brain MR image classification. -The technique is easy in implementation -Fast in execution and accurate in classification.	-The proposed methodology running time is very low, but it requires some huge memory to run.	-This proposed methodology achieved 99.38% accuracy. Whereas, the computation time is very low.
[8]	- MR images classification -The grayscale image is converted to RGB and calculate each channel of RGB and extracted 9	- Median filter -Color features -Mean, variance and skewness -k-NN	-The accuracy rate is better as compare to other methods. -The k-NN is efficient for the smaller dataset.	-The accuracy of k-NN gets reduces when the dataset is large.	-A total of 100 images has been used. -30% testing and 70% of training criteria have been used. -The accuracy rate for

	features using color moments				normal and abnormal images during training and testing is 98% and 95% respectively. -Low computation time for feature extraction and classification.
[9]	-Cerebrospinal fluid -Gray matter -White matter -Brain tumor	-Fuzzy filter -Type-II Probabilistic C-Mean -Type-II Fuzzy Logic -Thresholding -Fuzzy clustering	Type-II fuzzy showed more effectiveness than Type-I.	-The proposed methodology accuracy rate is low. -In the pre-processing stage, the Type-II fuzzy expert system needs further improvement.	-95 total images have been used. - 79 images were identified correctly and 16 were identified incorrectly and obtained a 78.94% accuracy rate.
[10]	Multi-resolution information is used to differentiate different tissues. -By combining stationary wavelet transform (SWT) coefficients and their statistical features, a multi-dimensional feature vector is formed.	-Anisotropic diffusion filter. -SWT. -Spatial filters -Self-Organization Map (SOM). -Learning Vector Quantization. -Internet brain segmentation repository	-As shown in the result of splitting texture information into different frequency channels, the SWT is very effective. -SOM works efficiently for dividing M x N-dimensional data into multiple segments.	-The execution time increases with the division of images in channels.	-The methodology is effective than manual segmentation, however, the accuracy is not mentioned in the paper.
[11]	Linear correlation coefficient -Linear regression -Four different classes of images such as cancer class, edema class, cancer class, and non-classified class.	-Principal component analysis -Gradient Descent with momentum weight and bias learning function -Artificial neural network -Feedforward artificial neural network	-The artificial neural network is simple and easy for classification. -The proposed method is fast in execution and easy in implementation and efficient in classification.	-For ANN the trainlm is used as a training function which is fast, however, it needs a lot of memory to run.	-The accuracy rate of the methodology is 96.33%. -The average consuming time is 0.2434s.
[12]	-Brain MRI segmentation -Images of different diseases have been considered such as normal, tumoral and MS. -GLCM calculates the co-occurrence matrix of each image. -Mean, variance and entropy features are extracted	-The gray level co-occurrence matrix -Principal component analysis -Linear discriminant analysis (LDA) -Artificial neural network (ANN) -k-nearest neighbor (k-NN)	-ANN is simple and also easy in implementation While for smaller dataset k-NN is more efficient -The accuracy rate of the proposed methodology is astounding compared to the existing methods.	-This technique may not effective for other modalities.	-Achieved a 100% accuracy rate for normal images, 92.86% for MS and 100% accuracy for tumoral images with ANN and k-NN classifiers. -The computation time for feature extraction and classification is only 0.025s.
[13]	-Color converted images -Differentiate the region and lesion size	-Thresholding -color converted segmentation -k-means clustering algorithm	-Fast and easy for small data -The accuracy rate of the methodology is quite high, whereas, the computation time is reasonable.	-k-means clustering is not effective for the choice of distance measure and the initial cluster assignment.	-100% accuracy rate for MRI T2-w images and the computation time is the 30s. - 75% accuracy rate for spin density and computation time is only 30s. -For T1-w images are 80% accuracy rate and 30s is the computation time.
[14]	-Detection of the mass tumor -Shape and range of brain tumor detection.	-Median filter -Euclidean distance -k-mean clustering -Fuzzy c-mean algorithm	-Fuzzy C-mean algorithm is accurate than k-means clustering	-The accuracy rate has not been shown -The noise addition and removal are not well explained.	-Results have not been shown clearly.
[15]	-Region-based global change. -2D MR slices -The region containing the tumor is considered	- The fast bounding box (FBB). -Score function. -Mean shift clustering (MSC). -Bhattacharya coefficient. -Ellipse fitting technique.	-Prior parameter distribution is not required for the unsupervised techniques. -Image registration is not	-The performance of FBB is mostly reduced by noise. -On MR slice this method generates a	-92% accuracy has been achieved for tumor detection and 89% for Edema by the proposed

	as a change in the image.		required for the fast bounding box. -Intensity standardization is not required in MR slices. -The labeled images training set is not required	box -The FBB performance depends on the asymmetry of two halves of MR images.	methodology.
[16]	-Eigen Vector. -Wavelet Coefficients -Asymmetry in an axial brain MR images	-Principal component analysis (PCA) -Discrete Wavelet Transform - Artificial neural network -Levenberg-Marquardt learning Rule.	-k-NN is efficient for the smaller dataset and ANN is simple in implementation -The complexity is removed by PCA. -Good accuracy rate is shown by the proposed methodology.	-From 1024 to 7 only, the features are reduced which is not adequate while the increase in features badly decreases the performance of the system.	k-NN achieved 99% accuracy rate. -ANN achieved 90% accuracy rate.
[17]	-Standard deviation -Seed point. -Different pixels intensity. -The arithmetic means of neighbor pixels.	-Arithmetic mean. -Median filter. -Standard deviation -Region growing algorithm.	-Due to the easiness of the method, the results are satisfactory. -The median filter preserves the image edges.	-The execution time gets increase with pre-processing and post-processing methods.	-The time of execution is less than 20s. -while the segmentation execution time is 2.9s.
[18]	-Image segmentation -Template based k-means and modified fuzzy c-means clustering (TKFCM) -Statistical feature and region-based feature.	-Adaptive threshold -Histogram imaging -wiener filter -median filter -TKFCM based segmentation. -Support vector machine -Artificial neural network back-propagation	-The SVM classifier handles both separable and non-separable problems. -First order statistic features used to determine brain normality and abnormality using SVM. -While ANN is simple in implementation and is used to reduce the error value. -ANN-BP is used to find the tumor regions.	-The accuracy rate of the proposed method is high but the use of many techniques leads to complexity.	-This method achieved 97.37% accuracy rate -Computation time is 2 min which is quite high. -sensitivity and specificity rate is 98% and 100% respectively.
[19]	-Image segmentation -Discrete Wavelet Transform (DWT) DAUB-4 -T2-W brain MR images -Level-4 decomposition is employed.	-Daubechies (DAUB-4) - Principal component analysis (PCA). -Support vector machine (SVM), Linear kernel and Radial basis kernel function (RBF).	-Two types of SVM kernel have been used. -The linear kernel does not perform well due to the small margin in feature space. -Radial basis kernel function performs well because support vectors automatically achieve the training process part.	-The low selection of features might lose important information. -only 7 features are selected per image. -DAUB-4 is computationally expensive.	-A total of 75 images of T2-W were used in which 10 were normal and 65 were abnormal images. - 98.46% accuracy has been achieved using SVM radial base function. -While, SVM linear base function achieved 94.7% accuracy.
[20]	-mBm texture features. -Probability distribution - Fractal dimension.	-KLD method -Fractal dimension, level set based method. -Expectation maximization algorithm. -Graph cut procedure. -Laplacian matrix.	-For tumor segmentation, this technique performs well. -The advantage of this method is, it can be implemented on various modalities.	-The time is very high for normalization, feature extraction, feature selection, and segmentation. -For different modalities, accuracy is not provided.	-In this methodology, 249 images of MRI have been considered. -100% accuracy has been shown by mBm features in multimodalities T1,T2 and FLAIR MRI segmentation.
[21]	Detection of volume changes in brain tissues	-Fuzzy C-Means Algorithm. -Artificial neural network -High pass and low pass filter -Mean and standard deviation.	-The number of iteration is decreased and the pixels are classified into one group.	-NN requires a huge amount of data and time for training. -Its computation is complex.	-The computation time is reasonable.
[22]	-Discrete wavelet transform (DWT) is used to extract wavelet coefficients -Gray level co-occurrence matrix is used for statistical feature extraction. -image segmentation.	-Morphological operation. -DWT -GLCM -Probabilistic neural network	-The method is easy in implementation, the accuracy rate is satisfactory. - The methodology of tumor detection and identification of location rate are quite speedy. -The probabilistic neural	-The computation time is high due to many proposed stages.	-100% accuracy rate is achieved for the training dataset -As from LL and HL, the statistical features were extracted. -While 95% accuracy rate is achieved for testing dataset.

			network quickly learns from training data and can adapt its learning in real-time.		
[23]	-Segmented White matter, gray matter, cerebrospinal fluid, and tumor tissues. -Gray level co-occurrence matrix. -Tumor detection.	-Berkeley wavelet transform (BWT) -Morphological operation. -Gray level Co-occurrence matrix (GLCM). -Support vector machine.	-The method is robust and obtained a good accuracy rate. -Significant for brain tumor detection. -Classification accuracy is not mentioned	- Support vector machine works effectively; however, the accuracy of SVM gets affected on small datasets.	-T1-W, T2-W and FLAIR images were used. -Total 96.51% accuracy was obtained as identifying normal and abnormal tissue from the MR images.
[71]	Deep neural network (DNN)	No need for pre-processing, feature extraction and feature reduction	Efficiently extracted the complex feature -Less outlier compare to other suggested methodologies	-Post-processing implementation is required	BRATS2013 dataset was used while the accuracy is not mentioned.
[72]	Based on fully CNN	Dice loss, bootstrapping loss, and sensitivity	Shows powerful and effective classification as compared to the original design of CNN	The ratio of false-positive prediction is high in image classification -Also, require high memory storage	-the accuracy rate is not mentioned
[73]	Pre-trained CNN	ResNet34	Shows powerful and effective classification	Require high memory storage	Achieved 100% accuracy
[74]	Stationary WT	SWT + GCNN	--	Post-processing is required	Obtained 98.6% accuracy
[75]	Tumor segmentation	Multi-cascade CNN and Conditional random fields	--	High memory storage is required	Obtained 88.24% accuracy

V. Discussion

We have observed from our literature review that most of the researchers focused on automatic methods. The median filter is extensively used in the pre-processing stage because the brain MRI images are mostly effected by salt and pepper noise. DWT and GLCM are the most used techniques for feature extraction. Principal component analysis and genetic algorithm are widely used for feature reduction. Lastly, the classification of brain MRI image task was mostly performed by using ANN, K-NN, SVM, and SOM. We have also observed from our literature review that CNN based algorithms provide some good results.

CNN acts differently from other classifiers. CNN learns its features automatically from an image, and there is no need to use the feature extraction or feature reduction stages to feed the features to CNN [70].

We have critically analyzed and reviewed different proposed MRI classification methods which pave the way for further research in this area which is the main contribution of this work.

VI. Conclusion and Future Work

The aim of this article is to retrospect the current trends in brain MRI classification. In this study, brain MR images have been used and proposed different methodologies by different researchers to classify the brain MR images into normal such as (gray matter, white matter, cerebrospinal fluid) and abnormal tissues such as (tumor infected tissues). Most researchers have used four stages to classify the brain MR image such as Pre-processing, Feature extraction, Feature reduction, and Classification. In pre-processing the noise is removed from the brain MR image by using a median filter mostly, as this filter removes the noise from MR image efficiently and preserves the edges effectively, as well as improving the quality of the image for further processing. The feature extraction is the second stage in which the important suitable features in the MR image are identified to detect the brain tumor. Most of the researchers have used a discrete wavelet transform (DWT) and Gray level co-occurrence matrix. The third stage is feature reduction, where the dimensionality of data is reduced to get the most favorable features from the image and in this stage mostly, principal component analysis has been used. The fourth and last stage is, classification of brain MR image as normal or abnormal and in this case, artificial neural network (ANN), k-nearest neighbor (k-NN) and support vector (SVM) have been used. In this paper, we have analyzed that still profound improvements are required in the segmentation and accuracy of MR image processing techniques. As we discussed in the literature that previous proposed techniques comprised of some serious shortcomings such as execution time and the accuracy of abnormality detection that needs to be solved. So, in this case, principal component analysis (PCA) which is a statistical technique used in data analysis is the best option to overcome these limitations.

VII. Conflict of Interest

The authors declare that they have no conflict of interest regarding the publication of this article.

Acknowledgments

We are grateful to Department of Computer Engineering, Changwon National University for the immense support in piloting this review. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2018R1C1B6008187).

REFERENCES

[1] E. S. A. El-Dahshan, T. Hosny, and A. B. M. Salem, “Hybrid intelligent techniques for MRI brain images classification,” Digit. Signal Process. A Rev. J., vol. 20, no. 2, pp. 433–441, 2010.

[2] M. Nazir, F. Wahid, and S. Ali Khan, “A simple and intelligent approach for brain MRI classification,” J. Intell. Fuzzy Syst., vol. 28, no. 3, pp. 1127–1135, 2015.

[3] R. Lavanyadevi, M. MacHakowsalya, J. Nivethitha, and A. Niranjil Kumar, “Brain tumor classification and segmentation in MRI images using PNN,” Proc. - 2017 IEEE Int. Conf. Electr. Instrum. Commun. Eng. ICEICE 2017, vol. 2017-Decem, pp. 1–6, 2017.

[4] J. Amin, M. Sharif, M. Yasmin, and S. L. Fernandes, “A distinctive approach in brain tumor detection and classification using MRI,” Pattern Recognit. Lett., vol. 0, pp. 1–10, 2017.

[5] S. Kumar, C. Dabas, and S. Godara, “Classification of Brain MRI Tumor Images: A Hybrid Approach,” Procedia Comput. Sci., vol. 122, pp. 510–517, 2017.

[6] Z. Iscan, Z. Dokur, and T. Ölmez, “Tumor detection by using Zernike moments on segmented magnetic resonance brain images,” Expert Syst. Appl., vol. 37, no. 3, pp. 2540–2549, 2010.

[7] Y.-D. Zhang and L. Wu, "An Mr Brain Images Classifier via Principal Component Analysis and Kernel Support Vector Machine," Progress In Electromagnetics Research, Vol. 130, pp. 369-388, 2012.

[8] M. Fayaz, A. S. Shah, F. Wahid, and A. Shah, “A Robust Technique of Brain MRI Classification using Color Features and K-Nearest Neighbors Algorithm,” Int. J. Signal Process. Image Process. Pattern Recognit., vol. 9, no. 10, pp. 11–20, 2016.

[9] M. H. Fazel Zarandi, M. Zarinbal, and M. Izadi, “Systematic image processing for diagnosing brain tumors: A Type-II fuzzy expert system approach,” Appl. Soft Comput. J., vol. 11, no. 1, pp. 285–294, 2011.

[10] A. Demirhan and I. Güler, “Combining stationary wavelet transform and self-organizing maps for brain MR image segmentation,” Eng. Appl. Artif. Intell., vol. 24, no. 2, pp. 358–367, 2011.

[11] W. H. Ibrahim, A. A. A. Osman, and Y. I. Mohamed, “MRI brain image classification using neural networks,” Proc. - 2013 Int. Conf. Comput. Electr. Electron. Eng. ’Research Makes a Differ. ICCEEE 2013, pp. 253–258, 2013.

[12] S. Jafarpour, Z. Sedghi, and M. C. Amirani, “A Robust Brain MRI Classification with GLCM Features,” Int. J. Comput. Appl., vol. 37, no. 12, pp. 975–8887, 2012.

[13] L. H. Juang and M. N. Wu, “MRI brain lesion image detection based on color-converted K-means clustering segmentation,” Meas. J. Int. Meas. Confed., vol. 43, no. 7, pp. 941–949, 2010.

[14] J. Selvakumar, A. Lakshmi, and T. Arivoli, “Brain tumor segmentation and its area calculation in brain MR images using K-mean clustering and Fuzzy C-mean algorithm,” IEEE-International Conf. Adv. Eng. Sci. Manag. ICAESM-2012, pp. 186–190, 2012.

[15] B. N. Saha, N. Ray, R. Greiner, A. Murtha, and H. Zhang, “Quick detection of brain tumors and edemas: A bounding box method using symmetry,” Comput. Med. Imaging Graph., vol. 36, no. 2, pp. 95–107, 2012.

[16] N. H. Rajini and R. Bhavani, “Classification of MRI brain images using k-nearest neighbor and artificial neural network,” Int. Conf. Recent Trends Inf. Technol. ICRTIT 2011, pp. 863–868, 2011.

[17] T. Węgliński and A. Fabijańska, “Brain Tumor Segmentation from MRI Data Sets Using Region Growing Approach.” In proc. of 2011 VIIth International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH), pp. 185-188, May 11-14, 2011.

[18] R. Ahmmed, A. Sen Swakshar, M. F. Hossain, and M. A. Rafiq, “Classification of tumors and it stages in brain MRI using support vector machine and artificial neural network,” ECCE 2017 - Int. Conf. Electr. Comput. Commun. Eng., pp. 229–234, 2017.

[19] M. Ahmad, “Classification of Tumors in Human Brain MRI using Wavelet and Support Vector Machine,” IOSR J. Comput. Eng., vol. 8, no. 2, pp. 25–31, 2012.

[20] S. Ahmed, K. M. Iftekharuddin, R. J. Ogg, and F. H. Laningham, “Efficacy of texture, shape, and intensity features for robust posterior-fossa tumor segmentation in MRI,” Med. Imaging 2009 Comput. Diagnosis, vol. 7260, no. February, p. 726020, 2009.

[21] K.J. Shanthi, M.N. Sasikumar and C. Kesavadas, “Neuro-fuzzy approach toward segmentation of brain MRI based on intensity and spatial distribution,” Jour. Of Med. Imaging and Radiation Sciences, Vol. 41, no. 2, pp. 66-71, 2010.

[22] N. Varuna Shree and T. N. R. Kumar, “Identification and classification of brain tumor MRI images with feature extraction using DWT and probabilistic neural network,” Brain Informatics, vol. 5, no. 1, pp. 23–30, 2018.

[23] N. B. Bahadure, A. K. Ray, and H. P. Thethi, “Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM,” Int. J. Biomed. Imaging, vol. 2017, 2017.

[24] D. N. Louis et al., “The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary,” Acta Neuropathol., vol. 131, no. 6, pp. 803–820, 2016.

[25] S. J. Choi, J. S. Kim, J. H. Kim, S. J. Oh, J. G. Lee, C. J. Kim, Y. S. Ra, J. S Yeo, J. S Ryu, D. H. Moon, “[18F] 3-deoxy-3-fluorothymidine PET for the diagnosis and grading of brain tumors,” European journal of nuclear medicine and molecular imaging, vol. 32, no. 6, pp. 653-659, 2005.

[26] M. Computing, “Image Segmentation for Tumor Detection,” vol. 2, no. May, pp. 244–248, 2013.

[27] B. Dong, A. Chien, and Z. Shen, “Frame based segmentation for medical images,” Commun. Math. Sci., vol. 9, no. 2, pp. 551–559, 2011.

[28] J. Patel and K. Doshi, “A Study of Segmentation Methods for Detection of Tumor in Brain MRI,” Adv. Electron. Electr. Eng., vol. 4, no. 3, pp. 279–284, 2014.

[29] R.C. Gonzalez, R.E. Woods, “Digital Image Processing, second ed., Prentice Hall, New Jersey,” 2002.

[30] A. Ortiz, J. M. Górriz, J. Ramírez, D. Salas-González, and J. M. Llamas-Elvira, “Two fully-unsupervised methods for MR brain image segmentation using SOM-based strategies,” Appl. Soft Comput. J., vol. 13, no. 5, pp. 2668–2682, 2013.

[31] I. N. Bankman, Handbook of Medical Imaging, Processing and Analysis, Academic Press, CA, 2000

[32] K. S. Thara and K. Jasmine, “Brain Tumour Detection in MRI Imagesusing PNN andGRNN,” 2016 Int. Conf. Wirel. Commun. Signal Process. Netw., pp. 1504–1510, 2016.

[33] M. G. Ballanger, E. Biglieri, S. Furui, Y. F. Huang, N. Jayant, A. K. Katsagglelos, M. Kaveh, P. K. R. Rajasekaran, J. A. Sorenson, Signal Processing and Communications, Taylor, & Francis, Boca Raton, 2005.

[34] E. Abdel-Maksoud, M. Elmogy, and R. Al-Awadi, “Brain tumor segmentation based on a hybrid clustering technique,” Egypt. Informatics J., vol. 16, no. 1, pp. 71–81, 2015.

[35] N. Nabizadeh and M. Kubat, “Brain tumors detection and segmentation in MR images: Gabor wavelet vs. statistical features,” Comput. Electr. Eng., vol. 45, pp. 286–301, 2015.

[36] M. Havaei, P. M. Jodoin, and H. Larochelle, “Efficient interactive brain tumor segmentation as within-brain kNN classification,” Proc. - Int. Conf. Pattern Recognit., pp. 556–561, 2014.

[37] N. Nabizadeh and M. Kubat, “Brain tumors detection and segmentation in MR images: Gabor wavelet vs. statistical features,” Comput. Electr. Eng., vol. 45, pp. 286–301, 2015.

[38] K. Somasundaram and T. Kalaiselvi, “Fully automatic brain extraction algorithm for axial T2-weighted magnetic resonance images,” Comput. Biol. Med., vol. 40, no. 10, pp. 811–822, 2010.

[39] W. Ayadi, W. Elhamzi, I. Charfi, and M. Atri, “A hybrid feature extraction approach for brain MRI classification based on Bag-of-words,” Biomed. Signal Process. Control, vol. 48, pp. 144–152, 2019.

[40] M. M. Subashini, S. K. Sahoo, V. Sunil, and S. Easwaran, “A non-invasive methodology for the grade identification of astrocytoma using image processing and artificial intelligence techniques,” Expert Syst. Appl., vol. 43, pp. 186–196, 2016.

[41] S. L. Jui et al., “Brain MRI Tumor Segmentation with 3D Intracranial Structure Deformation Features,” IEEE Intell. Syst., vol. 31, no. 2, pp. 66–76, 2016.

[42] E. A. S. El-Dahshan, H. M. Mohsen, K. Revett, and A. B. M. Salem, “Computer-aided diagnosis of human brain tumor through MRI: A survey and a new algorithm,” Expert Syst. Appl., vol. 41, no. 11, pp. 5526–5545, 2014.

[43] S. Chaplot, L. M. Patnaik, and N. R. Jagannathan, “Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network,” Biomed. Signal Process. Control, vol. 1, no. 1, pp. 86–92, 2006.

[44] P. Georgiadis et al., “Improving brain tumor characterization on MRI by probabilistic neural networks and non-linear transformation of textural features,” Comput. Methods Programs Biomed., vol. 89, no. 1, pp. 24–32, 2008.

[45] M. Havaei, P. M. Jodoin, and H. Larochelle, “Efficient interactive brain tumor segmentation as within-brain kNN classification,” Proc. - Int. Conf. Pattern Recognit., pp. 556–561, 2014.

[46] M. Sharma, and S. Mukharjee, “Brain Tumor Segmentation using hybrid Genetic Algorithm and Artificial Neural Network Fuzzy Inference System (ANFIS),” Int. J. Fuzzy Log. Syst., vol. 2, no. 4, pp. 31–42, 2012.

[47] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.M. Jodoin, H. Larochelle, “Brain tumor segmentation with Deep Neural Networks,” Med. Image Anal., vol. 35, pp. 18–31, 2017.

[48] S. Roy, S. Sadhu, S. K. Bandyopadhyay, D. Bhattacharyya, and T. H. Kim, “Brain tumor classification using adaptive neuro-fuzzy inference system from MRI,” Int. J. Bio-Science Bio-Technology, vol. 8, no. 3, pp. 203–218, 2016.

[49] K. A. Smitha, A. K. Gupta, and R. S. Jayasree, “Relative percentage signal intensity recovery of perfusion metrics-an efficient tool for differentiating grades of glioma,” Br. J. Radiol., vol. 88, no. 1052, 2015.

[50] Y. Zhang and L. Wu, “An MR brain images classifier via principal component analysis and kernel support vector machine,” Prog. Electromagn. Res., vol. 130, no. June, pp. 369–388, 2012.

[51] A. Işin, C. Direkoğlu, and M. Şah, “Review of MRI-based Brain Tumor Image Segmentation Using Deep Learning Methods,” Procedia Comput. Sci., vol. 102, no. August, pp. 317–324, 2016.

[52] T. Akila and P. UmaMaheswari, “Comparative analysis of classifier performance on MR brain images,” Int. Arab J. Inf. Technol., vol. 12, no. 6A, pp. 772–779, 2015.

[53] B. Sudha, P. Gopikannan, A. Shenbagarajan, C. Balasubramanian, “Computers and Software (IRECOS),” Int. Rev. Comput. Softw., vol. 9, no. 4, 2014.

[54] S. B.Gaikwad and M. S. Joshi, “Brain Tumor Classification using Principal Component Analysis and Probabilistic Neural Network,” Int. J. Comput. Appl., vol. 120, no. 3, pp. 5–9, 2015.

[55] N. Gordillo, E. Montseny, and P. Sobrevilla, “State of the art survey on MRI brain tumor segmentation,” Magn. Reson. Imaging, vol. 31, no. 8, pp. 1426–1438, 2013.

[56] E. A. S. El-Dahshan, H. M. Mohsen, K. Revett, and A. B. M. Salem, “Computer-aided diagnosis of human brain tumor through MRI: A survey and a new algorithm,” Expert Syst. Appl., vol. 41, no. 11, pp. 5526–5545, 2014.

[57] M. Fayaz, A. S. Shah, F. Wahid, and A. Shah, “A Robust Technique of Brain MRI Classification using Color Features and K-Nearest Neighbors Algorithm,” Int. J. Signal Process. Image Process. Pattern Recognit., vol. 9, no. 10, pp. 11–20, 2016.

[58] N. Elaiza, A. Khalid, S. Ibrahim, and U. T. Mara, “MRI Brain Abnormalities Segmentation using,” vol. 3, no. 2, pp. 980–990, 2011.

[59] C. J. C. Burges and B. Schölkopf, “Improving the accuracy and speed of support vector machine,” Adv. Neural Inf. Process. Syst., no. June, pp. 375–381, 1997.

[60] D. Decoste and B. Schölkopf, “Training invariant support vector machines,” Mach. Learn., vol. 46, no. 1–3, pp. 161–190, 2002.

[61] E. I. Zacharaki et al., “MRI-based classification of brain tumor type and grade using SVM-RFE,” Proc. - 2009 IEEE Int. Symp. Biomed. Imaging From Nano to Macro, ISBI 2009, pp. 1035–1038, 2009.

[62] H. Selvaraj, S. Thamarai Selvi, D. Selvathi & L. Gewali, “Brain MRI Slices Classification Using Least Squares Support Vector Machine,” International Journal of Intelligent Computing in Medical Sciences & Image Processing, 1:1, pp. 21-33, 2007.

[63] Ahmed Kharrat, Karim Gasmi, Mohamed Ben Messaoud, Nacéra Benamrane, and and Mohamed Abid, “A Hybrid approach for automatic classification of brain MRI using genetic algorithm and support vector machine,” Leonardo J. Sci., vol. 9, no. 17, pp. 71–82, 2010.

[64] S. Sindhumol and K. Balakrishnan, “Abnormality Detection from Multispectral Brain MRI using Multiresolution Independent Component Analysis,” vol. 6, no. 1, pp. 177–190, 2013.

[65] V. S. Takate and P. S. Vikhe, “Classification of MRI Brain Images using K-NN and k-means,” no. 2, pp. 55–58, 2012.

[66] G. Vishnuvarthanan, M. P. Rajasekaran, P. Subbaraj, and A. Vishnuvarthanan, “An unsupervised learning method with a clustering approach for tumor identification and tissue segmentation in magnetic resonance brain images,” Appl. Soft Comput. J., vol. 38, pp. 190–212, 2016.

[67] N. Dhanachandra, K. Manglem, and Y. J. Chanu, “Image Segmentation Using K-means Clustering Algorithm and Subtractive Clustering Algorithm,” Procedia Comput. Sci., vol. 54, pp. 764–771, 2015.

[68] M. N. Wu, C. C. Lin, and C. C. Chang, “Brain tumor detection using color-based K-means clustering segmentation,” Proc. - 3rd Int. Conf. Intell. Inf. Hiding Multimed. Signal Process. IIHMSP 2007., vol. 2, no. December 2007, pp. 245–248, 2007.

[69] L. H. Juang and M. N. Wu, “MRI brain lesion image detection based on color-converted K-means clustering segmentation,” Meas. J. Int. Meas. Confed., vol. 43, no. 7, pp. 941–949, 2010.

[70] Y. Pan et al., “Brain tumor grading based on Neural Networks and Convolutional Neural Networks,” Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2015-November, pp. 699–702, 2015.

[71] M. Havaei et al., “Brain tumor segmentation with Deep Neural Networks,” Med. Image Anal., vol. 35, pp. 18–31, 2017.

[72] J. Zhang, X. Shen, T. Zhuo, and H. Zhou, “Brain Tumor Segmentation Based on Refined Fully Convolutional Neural Networks with A Hierarchical Dice Loss,” 2017.

[73] M. Talo, U. B Baloglu, Ö. Yıldırım, U. Rajendra Acharya, “Application of deep transfer learning for automated brain abnormality classification using MR images,” Cognitive Systems Research, 54, pp. 176-188, 2019.

[74] M. Mittal, L. Mohan Goyal, S. Kaur, I. Kaur, Amit Verma, D. Jude Hemanth, “Deep learning based enhanced tumor segmentation approach for MR brain images,” Applied Soft Computing, 78, pp. 346-354, 2019.

[75] K. Hu, Q. Gan,Y. Zhang, S. Deng, F. Xiao, W. Huang, C. Cao and X. Gao, “Brain Tumor Segmentation Using MultiCascaded Convolutional Neural Networks and Conditional Random Field,” IEEE Access,7, pp. 92615-92629, 2019.

[76] T. M. Mohammed, K. K. Pradeep Mohan, Murugan, B. S., Dhanasekeran, S., Vijayakarthick, P., Selvi, A. Senthil, "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images," Journal of Medical Systems, 43 (9), pp. 294, 2019.

[77] K. Hu, Q. Gan,Y. Zhang, S. Deng, F. Xiao, W. Huang, C. Cao and X. Gao, "Brain Tumor Segmentation Using Multi Cascaded Convolutional Neural Networks and Conditional Random Field," IEEE Access,7, pp. 92615-92629. 2019.

[78] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, Y. Fan, "A deep learning model integrating FCNNs and CRFs for brain tumor segmentation," Medical Image Analysis, 43, (2018), pp. 98-111.

[79] T. Chithambaram and K. Perumal, Brain tumor segmentation using genetic algorithm and ANN techniques, IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, pp. 970-982, 2017.

[80] J. Li, Z. Liang Yu, Z. Gu, H. Liu, Y. Li, "MMAN: Multi-modality aggregation network for brain segmentation from MR images," Neurocomputing, 358, pp. 10-19, 2019.

[81] K. Kamnitsas, C. Ledig, V. F.J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," Medical Image Analysis, 36, pp. 61-78, 2017.

[82] E. Abdel-Maksoud, M. Elmogy, R. Al-Awadi, "Brain tumor segmentation based on a hybrid clustering technique," Egyptian Informatics Journal,16 (1), pp. 71-81, 2015.

AUTHOR'S PROFILE

Zahid Ullah, is currently doing Ph.D in Computer Engineering from Changwon National University South Korea. He received MS in Computer Science from SZABIST, Islamabad, Pakistan in 2015, and BS in Information Technology from University of Malakand Pakistan in 2011. His area of interest are Image Processing and Machine Learning Techniques.

Dr. Su-Hyun, Lee, is working as Professor and Head of department of Computer Engineering, at Changwon National University South Korea. He is working in multiple desciplines, especially Bio-Informatics, Algorithm, and Programming.

Dr. Donghyeok An received the B.S. degree in computer science from Handong University, Pohang, South Korea, in 2006, and the Pd. D. degree in Computer Science from Korea Advanced Institute of Science and Technology (KAIST) in Feb. 2013. He was a post doctor and a visiting professor at Sungkyunkwan University from Mar. 2013 to Feb. 2014. He was a senior engineer at Samsung Electronics from 2014 to 2015. He was an assistant professor in the Department of Computer Engineering, Keimyung University from 2015 to 2017. He is currently an assistant professor in the Department of Computer Engineering, Changwon National University, Changwon, Korea. His research interests include 5G networks, ultra reliable low latency communication, and Internet of Things.

Secure V2V Communication in IOV using IBE and PKI based Hybrid Approach

Satya Sandeep Kanumalli[1]
Research Scholar, CSE Department
University College of Engineering
Acharya Nagarjuna University.and Asst.Professor
Vignan's Nirula Institute of Technology and Science for Women, Guntur 522009
India

Anuradha Ch[2]
CSE Department, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada 52007, India

Dr. Patanala Sri Rama Chandra Murty[3]
CSE Department, Acharya Nagarjuna University
Guntur 522510, India

***Abstract*—We live in the world of "Internet of Everything", which lead to advent of different applications and Internet of vehicles (IOV) of one among them, which is a major step forward for the future of transportation system. Vehicle to vehicle (V2V) communication plays a major role in which a vehicle may send sensitive, non-sensitive messages and these messages are encrypted with public keys, which makes distribution of public keys is a major problem due to the vehicle need to be anonymous having pseudonyms which changes more frequently and makes it more complicated. Here we proposed a hybrid approach, which uses existing Public key certificate for authorization of the vehicle and Identity Based Encryption to generate public keys from the pseudonyms and use it in secure V2V communication without compromising anonymity of the vehicle.**

Keywords—Privacy; internet of vehicles; hashing; IBE; public key certificate

I. Introduction

Due to advent of wireless communication and internet there are many things that are connected and communicate with each other with out the intervision of human and many applications are derived in which Internet of vehicles (IOV) stands tall as its is big thing which enables autonomous driving which changes the future of transportation [1][3].

IOV[17][18] is derived from age old VANETS in which the vehicles and its infrastructure are interconnected using internet technologies like 4G and 5G, but the most of the arcitecture remains the same as the treditional VANETs, which inherets some of the challenges like security and privicy of the vehicle and comminication as well.

Thre are different types of communication players Like Vehicle, RSU, CA etc and they can enable different types of communication like V2V, V2R etc and can also extand as V2X which enables the communication between the vehicle and pedestrian, smart objects etc. out of which V2V communication playes amajor role through the vehicle exchanges different types of messages like safety and non safety messages[15]. V2V communication playes amajor role in the whole IOV communication, through the vehicle exchanges different types of messages like safety and non safety messages, when it comes to sefity messages time is very critical and for non safety message time may not be critical[21].

V2V messages should be secured and yet not compramising the privicy of the vehicle, as vehicles uses masking identites called Pseudonyms. A vehicle communicates 100's of messages every ms and they are different forms of attacks an attacker may plan like attacks [10] on privacy compromising the location and the real identity of the vehicle of the vehicle, attacks on integrity of the message in which the attacker may modify the message, eavesdropping through which the attacker tries to read the communication, different solutions are been proposed to combat security problems in V2V communication, which are broadly classified in to three types.

a) Hardware based solutions in which the attacks are identified at the physical level[25][26] by utilizing channelization and using provability distribution functions in which these solutions only work with the attacks on physical layer and completely ignore the attacks on higher layers

b) IBE based solutions in which different solutions utilizes a trusted authority[8][9], which distributes the keys to the vehicles and some of the solutions utilizes certificates and some may not very time the trusted authority [16] must act as a middle man which incurs delay and bottle neck in the network. Some completely ignore the usage of pseudonyms, which compromises the privacy of vehicle.

c) The other solutions are based on trust-based mechanisms[22][23][24], in which every vehicle is associated with a trust value and may vary depending on vehicle behavior [19] and there also exist a trusted third party (TTP) or mutual group trust management, which evaluates and isolates the attacker depending on its trust value[27]. These solutions may not work all the time as the attacker may act normal to improve its trust value and the TTP may compromise.

In this paper we address the problem using a hybrid mechanism which utilizes the existing public key certificate and Identity Based Encryption (IBE) by which the vehicle can use pseudonyms and hide its identity and yet traceable by the Transport Authorities which can revoke the certificate in case of suspicious vehicle [28].

The rest of the paper organized as follows; in Section II we briefly discuss the existing approaches along with its drawbacks, in Section III we have presented the System Architecture we followed, in Section IV we thoroughly discussed our hybrid approach and also we presented the proof of work for IBE, in Section V we presented the threat model with the adversary and its different forms of attacks, in Section VI shows our simulation results, in Section VII presents conclusion and future work.

II. Related Work

Works related to physical layer security focus on physical aspects. Hanet al. [1],based on securing sub carrier allocation where the eavesdropper may intercept the communication, so the sub carrier allocation, joint rally selection is done by using the provability calculated by forming RGB, Random by parity Graph which given a good results as the work is based on physical layer security, where most of the attacks are undertaken on higher layers.

Jinyuan Sun et al. [20] Based on IBE which uses threshold based secret sharing and using pseudonyms for preserving privacy here RTA acts as middle man which generates the keys and it is susceptible to sing point of failure which make the whole system prone to DOS attack.

Liu, Yanbing et al. [3] have proposed an authentication and key agreement scheme for safeguarding V2V communication which involves third party Trusted authority in the communication which incurs delay as every time TA is also a part of authentication.

Debiao He et al. [4] proposed a scheme based on IBE which does not use Bilinear pairings, called CCPA conditional Privacy Preserving Authentication scheme which is better proved to reduce the computational cost which again shows the trusted authority plays a key role which ultimately be a single point of failure and the communication is made through internet which can be a bottle neck in the communication.

Song, Jun et al. [5] have proposed a IBE scheme, which is light weight and doesn't use certificates for authentication and RTA generates the master key and it also manages the communication without RSU, without public key certificates after fraud detection revocation is an issue.

T.W. Chim et al.[6] have proposed a privacy preservation schemes for unicast and group communication in which RSU stands as a middle man, which verifies the signatures which shows the reduction of message overhead, but it incurs load on RSU, and it may prone to Sybil attack.

Lee et al. [7] have proposed a batch message verification scheme based on bilinear pairing, message singing and verification process, is a bit complicated which incurs overhead in communication.

III. System Architecture

IOV has no standardized architecture defined, different types of architecture's proposed cloud based, fog and cloud hybrid architecture and these derived from basic VANET architecture.

Here we follow very basic IOV architecture shown in Fig. 1, which consists of vehicles V, equipped with Dedicated Short range communication DSRC [11]boards, these boards are fixed as a Tamper proof Device TPD [12], and these are called on board units OBU. DSRC is the modified form of standardWi-Fi called 802.11p [29], which uses 5.9 GHz band and the layers of 1609, which together called WAVE.

Roads, parking lots and other places in which a car can move are covered by Road Side Units (RSU), generally these are equipped with high computational power processors and wired and wireless technologies [30] like 4G, 5G, WAVE etc. RSUs are aligned in such a way that they are in a line of site to one another,for local and global handover management, and they are interconnected with RSU controller RSUC which acts as switching station and it also moves the data from TA and CA back and fourth.

As shown in Fig. 1 Transport Authority TA is the root of the system, which registers the vehicles soon after it is been purchased from the show room and it creates an entry in its database, which contains all the vehicle information, as well as driver information and other license data.

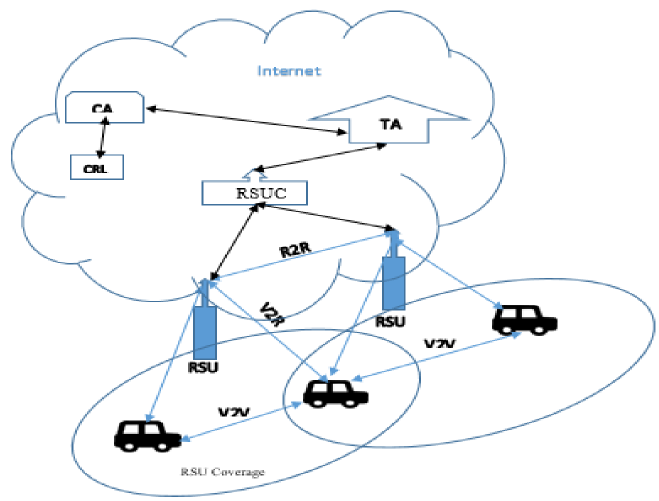


Fig. 1. IOV System Architecture.

Certification Authority CA Creates a Public key certificate for every vehicle after registration with TA, which contains public key information [31] along with the expiry and this information, can be accessible to RAUs as well and CA has a database CRL certificate revocation list [32] which can be used by RSU to send revocation request for vehicles found to be attackers[33].

IV. Proposed IBE-PKI based Hybrid Approach

The proposed system is based on bilinear maps on abelian groups based on [2][21], which maps the elements from two Additive groups G_1, G_2 to a target group G_T, here we follow a hybrid approach rather than pure IBE in which every vehicle should process the public key certificate issued by CA and on the top of it uses IBE to avoid the delay caused by certificate verification every time. The terminologies used in the manuscript are represented in Table I.

TABLE. I. TERMINOLOGIES USED

Abbreviation	Full form or meaning
V	Vehicle
V_A	Adversary
OBU	On board unit
TPD	Tamper Proof Device
WAVE	Wireless Access in Vehicular Networks
DSRC	Dedicated Short Range Communication
RSU	Road Side Units
V2V	Communication between vehicles
V2R	Communication between vehicle and RSU
R2R	Communication between RSUs
RSUC	RSU controller
TA	Transport Authority
CA	Certification Authority
CRL	Certificate revocation List
KCG	Key Generation Center
P	Prime Number
G, G_T	Generator and target Groups
MS_R	RSUs Master Secret
e	Bilinear map
H_1	Hash function maps $[0,1]^l$ -> G
H_2	Hash function maps G-> $[0,1]^l$
P_{id}	Pseudonym Identity
SK_{id}	Secret generated to P_{id}
M_{v2}	Message for Vehicle V_1 from V_2 in plain text
CT_{V2}	Cipher text generated for M_{v2}
||	Concatenation
$\oplus$	XOR operation
r	Random number in Z_p
Q	g^{α}

The proposed method can be viewed in five stages: i) Vehicle registration with Certification Authority is, ii) RSU Setup, iii) Vehicle registration with RSU, iv) Encryption by vehicle V_2, v) Decryption by Vehicle V_1 and some of these steps are carried only once in the communication scenario.

1) Vehicle registration with certification authority: Vehicle V soon after its onboard unit (OBU) is configured, public, private key pairs PUV, PRV are generated and sent along with its vehicle details, and unique identifier to Certification Authority (CA) for its public key certificate, CA verifies the details as shown in Fig. 2, store them to its database and issues the vehicle with its public key certificate CV. This step is carried only once in the vehicles lifetime unless the hardware gets changed or the Certificate is expired.

2) RSU setup: RSU are assumed to be with in the line of sight to one another and every time a vehicle moves from one RSU to another RSU, every RSU configures itself and generates master secret on hourly or daily basis.

a) RSU acts as Key Generation Center (KGC), it chooses a master secret MS_R and identifies global parameters perm (P, G, G_T, g, e)

P is a prime number

G, G_T are two cyclic groups generator and target, one additive and one multiplicative groups

g is the generator of G

e is the bilinear map element from G -> G_T

b) perm can also be given as (P, Q, H_1, H_2)

Where H_1 and H_2 are the hash functions

H_1 maps a string or id of the vehicle to element in G

$H_1 = [0,1]^l$ -> G

H_2 maps element in G_T back to the string

H_2 = G -> $[0,1]^l$

$MS_R = \alpha$

3) Vehicle registration with RSU: Vehicle holds the public key certificate, as it should not disclose its identity to the other vehicles it should not publically announce its certificate instead it gains pseudonyms by showing certificate to RSU as shown in Fig. 3. These pseudonyms can be changed when required or following pattern or randomly making the vehicle anonymous.

a) Vehicle V1 when it enters new RSU region, sends its public key certificate to RSU, RSU verifies the certificate and its period of validity and issues a set of pseudonyms (P_{id1}, P_{id2}, …P_{idn})$_{V1}$ which can be utilized for communication.

b) Vehicle V1 when it intends to change its pseudonym, it sends its new P_{id} to RSU, RSU generates secret key SK_{id} by applying H1 as follows and its master secret key α.

$SK_{id} = (H_1(P_{id}))^{\alpha}$

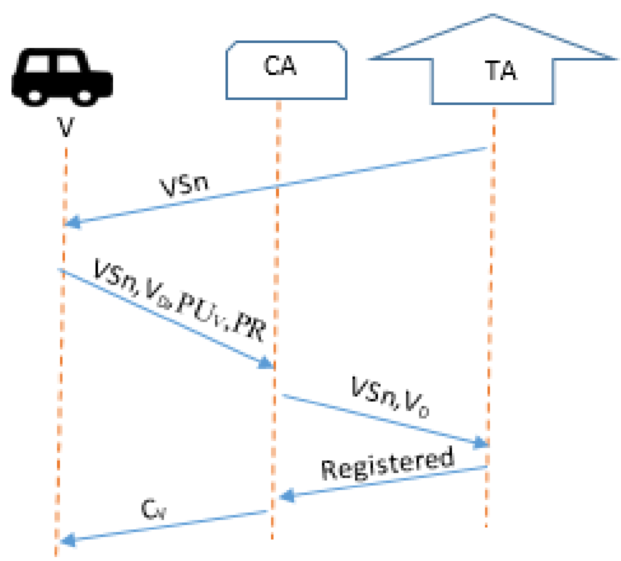


Fig. 2. Vehicle Registration with CA.

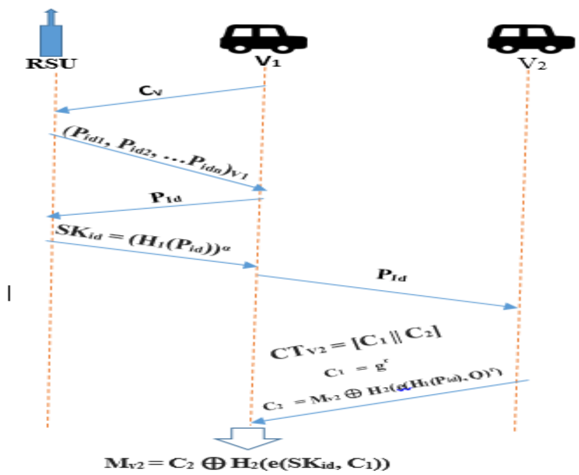


Fig. 3. Message Encryption and Decryption.

The above step maps the pseudonym which is string of $[0,1]^l$ to element in G^{α}

4) *Encryption by vehicle V_2:* When a vehicle tries to send a message may be Basic safety or emergency Broad Cast messages some of the messages needs encryption in certain cases and these messages must be confidential and as well as the system must preserve integrity so encryption is the only means.

a) If a vehicle V_2 wants to send a confidential message to vehicle V_1, V_2 prepares the message M_{v2} which is also a string of $[0,1]^l$.

b) It have its global parameters which is acquired from RSU (P, g, Q, H_1, H_2).

$Q = g^{\alpha}$

c) V_2 encrypts the message using pseudonymP_{id} of vehicle V_1, which produces the cypher text as follows.

$CT_{V2} = [C_1 \| C_2]$

$C_1 = g^r$

r is the random number in Z_p

$$C_2 = M_{v2} \oplus H_2(e(H_1(P_{id}), Q)^r) \quad (1)$$

d) Vehicle V_2 sends cipher text CT_{V2} to V_1

5) Decryption by Vehicle V_1

a) Vehicle V_2 soon after it receives CT_{V2} from V_2 it decrypt the message using its secret key SK_{id} as follows

$$M_{v2} = C_2 \oplus H_2(e(SK_{id}, C_1)) \quad (2)$$

b) Proof of correctness

$M_{v2} = C_2 \oplus H_2(e(SK_{id}, C_1))$

Substituting C_2 from eq(1) in eq(2)

$$= M_{v2} \oplus H_2(e(H_1(P_{id}), Q)^r) \oplus H_2(e(SK_{id}, C_1)) \quad (3)$$

Substituting SK_{id}, C_1 in eq (3)

$= M_{v2} \oplus H_2(e(H_1(P_{id}), Q)^r) \oplus H_2(e((H_1(P_{id}))^{\alpha}, g^r))$

Substituting Q in above equation

$= M_{v2} \oplus H_2(e(H_1(P_{id}), g^{\alpha})^r) \oplus H_2(e((H_1(P_{id}))^{\alpha}, g^r))$

$= M_{v2} \oplus H_2(e(H_1(P_{id}), g))^{\alpha r} \oplus H_2(e((H_1(P_{id})), g))^{\alpha r}$

$= M_{v2}$

V. Threat Model

Here we assume a Adversary which is planted at every RSU and it is having all the computation and communication technology, which also act as any other vehicle but the main job is to intercept the traffic, modify the messages, track the vehicles to map the original identities.

a) Attacks on confidentiality: When vehicle V_2 wants to send a message M to V_1, it encrypts the message and sends $CT_{V2} = [C_1 \| C_2]$ to V_1 and the adversary V_A tries to decrypt the message and to extract r out of C_1 is a complex and computing SK_{id} from the known P_{id} is a discrete logarithmic problem which V_A is incapable of computing.

b) Attacks on integrity: Vehicle V_2 sends cipher text in the form $[C_1 \| C_2]$ to V_1 and before it reaches V_1, Adversary V_A captures the message and tries to modify the message by preparing M_{v2} and prepare $C_2^{\wedge} = M_{v2} \oplus H_2(e(H_1(P_{id}), Q)^r)$ by choosing a random $r^{\wedge}$ and precomputing $C_1^{\wedge} = g^{r^{\wedge}}$ and sends $[C_1^{\wedge} \| C_2^{\wedge}]$ to V_1,.

V_1 computes $M_{v2}^{\wedge} = C_2^{\wedge} \oplus H_2(e(SK_{id}, C_1^{\wedge}))$

Decryption was a success and V_1takes the message as granted, to overcome this problem V_2 must also concatenate an encrypted hash along with the cipher text.

$CT_{V2} = [C_1 \| C_2 \| EH(M_{V2})]$

$EH(M_{V2})$ is the encrypted hash of M_{V2} with P_{id} of V_1 so that only V_1 can decrypt the hash and can check for integrity. As V_A can only tamper Message but not hash integrity of the message is preserved.

c) Attacks on anonymity: Anonymity should be preserved in IOV as the adversary can track the vehicle and understand its driving pattern, behavioral patterns, and may employee some physical attacks like kidnap, Murder, extortion, etc. For the adversary VA to get the real identity of the vehicle Vid it must get the certificate CV1 knowing Pid, only RSU is having the data associated with CV1 and we assume RSU to be a tamper proof device. It is impossible to the adversary.

As the pseudonyms keeps changing very time and the vehicle acquires new pseudonyms under new RSU and it is computationally infeasible to track the original identity of the vehicle from Pseudonyms.

VI. Simulation Results

A. Simulation Setup

For Simulation we make use of of a highway Junction road map at Guntur using open street map (OSM) and imported the map to SUMO [13] traffic simulator. We have created traffic nodes and other boundaries using polyconvert and netconvert commands contained in the SUMO 0.25.0, we have used randomTrips.py to create random trips for the vehicles with different intervals with variable arrival rates say

1, 1.2, 2, 2.5 vehicles every second each vehicle arrive into the map and depart at variable rates and we setup the speed range to a max of 60 kmph.

For network simulation, we use OMNET++ 5 [14], having Venius Package, which enables the network simulation for SUMO traffic having RSU and the nodes enabled with DSRC stacks, and we have created different communication scenarios with incremental number of vehicles and variable no of attackers, the attackers as discussed in the threat model have the same capabilities as the Vehicle node. These attackers are intentionally implanted in to the network at different points, and have two basic functionalities. Some of them defined to be in the radio coverage of a vehicle and listen to the traffic between RSU and Vehicle node, store the pseudonyms and try to track the vehicle even pseudonym keeps changing and the other are defined to eavesdropping the communication between the vehicles and try to decrypt the messages.

B. Performance Evaluation

Here we consider three scenarios in which we evaluate the performance of the system one is the 'idle–IOV' which evaluates the idle condition with no attackers and no security mechanism implemented, one called 'attack-IOV' in which we implant attackers and with no security mechanism implemented and the other is our proposed IBE-PKI based Hybrid approach called 'IPH-IOV', in which we implement our defense mechanism along with the attackers implanted.

C. Computational Overhead

It is the extra work done by the CPU to execute the whole scenario and it is calculated by summing up the CPU cycles consumed by all the nodes in the network.

Here on the x-axis, we plot the No. of vehicles varies from 0 to 250 and on Y-axis we took computational overhead % varies from 0 to 1 and Fig. 4 shows computational overhead in Idle-IOV is very low as there is no much additional computation done by the vehicles or by the system, In our proposed IBE-PKI the computational overhead is initially not much high but as the number of vehicles increases in the system, the number of computations also increases due to the key computations and exchanges in the system and the Attack-IOV is between Idle-IOV and IBE-PKI as the computation is not much needed and the attackers computation makes it deviated from the Idle-IOV.

D. End to End Delay

It is the total delay imposed on the packet from sender to the receiver and End-to-End delay is given as follows.

End to End Delay = Σ (packet arrival time – Packet send time) / Σ (No. of connections).

On x- axis, we plot the simulation time that varies from 10 to 50 sec and on y-axis, we plot End to End Delay in mille seconds, varying from 0 to 700 ms. Fig. 5 shows End to End delay of the Idle-IOV is very low compared to other two as there are no attackers in the system no packets are delayed and it increases with simulation time, on the other hand Attack-IOV have high End to End delay in the system due to the attackers delay or drop the packets.

Our IBE-PKI have shown some delay compared to Idle-IOV due to key generation, encryption and decryption, it lies very close to Idle-IOV and it gave a good result compared to Attack-IOV which is much higher.

E. Packet Delivery Ratio

It is the ratio of the packets sent and received

Packet Delivery ratio = Σ No. of packets send / Σ No. of packets received

On x- axis, we plot the No. of vehicles that varies from 40 to 200 and on y-axis we plot End to End Delay % that varies from 0 to 2.

Fig. 6 shows Packet Delivery Ratio of the Idle-IOV is very high compared to other two as there are no attackers in the system no packets are delayed or dropped and it increases with the increase in number of vehicles, on the other hand Attack-IOV have low Packet Delivery Ratio in the system due to the attackers delay or drop the packets which ultimately decreases Packet Delivery Ratio.

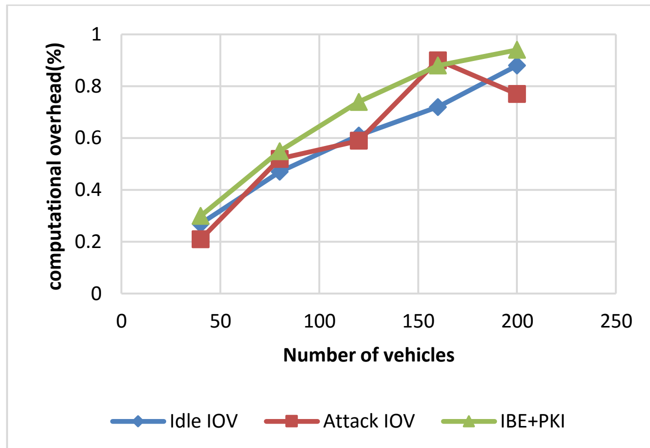


Fig. 4. Computational Overhead.

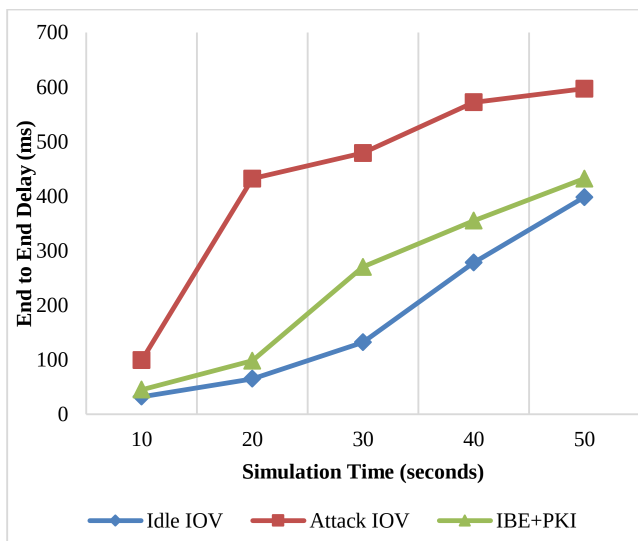


Fig. 5. End-to-End Delay.

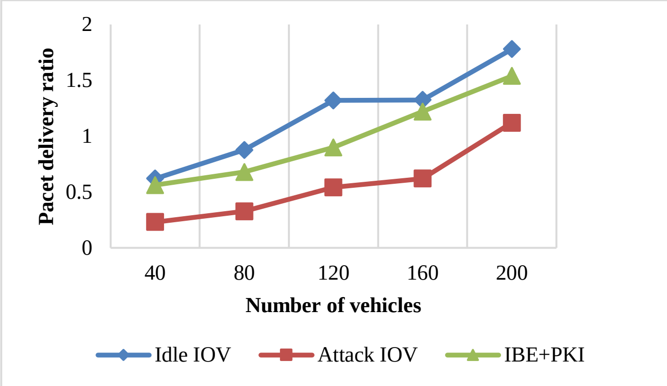


Fig. 6. Packet Delivery Ratio.

Our IBE-PKI have shown a very good result compared to Attack-IOV as it reduces the delay or drop of packets caused by the attackers and the results are very close to Idle-IOV.

F. Network Throughput

It is the sum of the packets delivered to the receivers successfully. Throughput is depicted in Fig. 7.

Throughput = Σ No. of packets received

On x- axis we plot the number of vehicle that varies from 40 to 200 and on y-axis, we plot Throughput in Megabits/ sec, varying from 0 to 1400 Mbps. Fig. 6 shows Network Throughput of the Idle-IOV is very high compared to other two as there are no attackers in the system no packets are delayed or dropped and it increases with the increase in number of vehicles, on the other hand Attack-IOV have low Network Throughput in the system due to the attackers delay or drop the packets which ultimately effects the Network Throughput value. Our IBE-PKI have shown a very good result compared to Attack-IOV as it reduces the delay or drop of packets caused by the attackers which increases the Network Throughput and the results are very close to Idle-IOV.

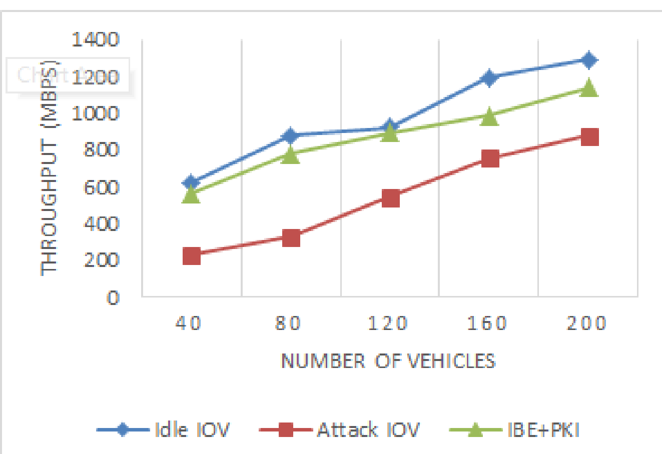


Fig. 7. Network Throughput.

VII. Conclusion and Future Work

Our Hybrid framework have safeguarded the system from the attackers yet not compromising the privacy of the vehicles and having significantly less computational delay and gave a good results compared to the other systems and the results obtained are close to the idle scenario. As a future work we try to still reduce the computational delay and also test the proposed mechanism in the real-world.

References

[1] Han, Di, Bo Bai, and Wei Chen. "Secure V2V communications via relays: Resource allocation and performance analysis." IEEE Wireless Communications Letters 6.3 (2017): 342-345.

[2] Boneh, Dan, and Matt Franklin. "Identity-based encryption from the Weil pairing." Annual international cryptology conference. Springer, Berlin, Heidelberg, 2001.

[3] Liu, Yanbing, Yuhang Wang, and Guanghui Chang. "Efficient privacy-preserving dual authentication and key agreement scheme for secure V2V communications in an IoV paradigm." IEEE Transactions on Intelligent Transportation Systems 18.10 (2017): 2740-2749.

[4] He, Debiao, et al. "An efficient identity-based conditional privacy-preserving authentication scheme for vehicular ad hoc networks." IEEE Transactions on Information Forensics and Security 10.12 (2015): 2681-2691.

[5] Song, Jun, et al. "Toward an RSU-unavailable lightweight certificate less key agreement scheme for VANETs." China Communications 11.9 (2014): 93-103.

[6] Chim, Tat Wing, et al. "SPECS: Secure and privacy enhancing communications schemes for VANETs." Ad Hoc Networks 9.2 (2011): 189-203.

[7] Lee, Cheng-Chi, and Yan-Ming Lai. "Toward a secure batch verification with group testing for VANET." Wireless networks 19.6 (2013): 1441-1449.

[8] S. Tangade, and S. S. Manvi, "Scalable and privacy-preserving authentication protocol for secure vehicular communications," in Proc. IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), 2017, pp. 1-6.

[9] Cui, Hui, Robert H. Deng, and Guilin Wang. "An Attribute-Based Framework for Secure Communications in Vehicular Ad Hoc Networks." IEEE/ACM Transactions on Networking(2019).

[10] Azees, Maria, Pandi Vijayakumar, and Lazarus Jegatha Deborah. "Comprehensive survey on security services in vehicular ad-hoc networks." IET Intelligent Transport Systems10.6 (2016): 379-388.

[11] Kenney, John B. "Dedicated short-range communications (DSRC) standards in the United States." Proceedings of the IEEE 99.7 (2011): 1162-1182.

[12] Mir, Zeeshan Hameed, and Fethi Filali. "LTE and IEEE 802.11 p for vehicular networking: a performance evaluation." EURASIP Journal on Wireless Communications and Networking 2014.1 (2014): 89.

[13] Krajzewicz, Daniel, et al. "SUMO (Simulation of Urban MObility)-an open-source traffic simulation." Proceedings of the 4th middle East Symposium on Simulation and Modelling (MESM20002). 2002.

[14] Varga, Andras. "OMNeT++." Modeling and tools for network simulation. Springer, Berlin, Heidelberg, 2010. 35-59.

[15] Kanumalli, Satya Sandeep, Anuradha Ch, and Patanala Sri Rama Chandra Murty. "Advances in Modelling and Analysis B." Journal homepage:http://iieta.org/Journals/AMA/AMA_B 61.1 (2018): 5-8.

[16] Kanumalli, Satya Sandeep, Anuradha Chinta, and Patanala Sri Rama Chandra Murty. "Isolation of Wormhole Attackers in IOV Using WPWP Packet Isolation of Wormhole Attackers in IOV Using WPWP Packet."

[17] Gerla, Mario, et al. "Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds." 2014 IEEE world forum on internet of things (WF-IoT). IEEE, 2014.

[18] Alam, Kazi Masudul, Mukesh Saini, and Abdulmotaleb El Saddik. "Toward social internet of vehicles: Concept, architecture, and applications." IEEE access 3 (2015): 343-357.

[19] Kanumalli, Satya Sandeep, Anuradha Ch, and Patanala Sri Rama Chandra Murty. "Advances in Modelling and Analysis B." Journal homepage:http://iieta.org/Journals/AMA/AMA_B 61.1 (2018): 5-8.

[20] Sun, Jinyuan, et al. "An identity-based security system for user privacy in vehicular ad hoc networks." IEEE Transactions on Parallel and Distributed Systems 21.9 (2010): 1227-1239.

[21] Lakshman Narayana Vejendla and Bharathi C R ,(2018),"Multi-mode Routing Algorithm with Cryptographic Techniques and Reduction of Packet Drop using 2ACK scheme in MANETs", Smart Intelligent Computing and Applications, Vo1.1, pp.649-658.DOI: 10.1007/978-981-13-1921-1_63.

[22] Lakshman Narayana Vejendla and Bharathi C R, (2018), "Effective multi-mode routing mechanism with master-slave technique and reduction of packet droppings using 2-ACK scheme in MANETS", Modelling, Measurement and Control A, Vol.91, Issue.2, pp.73-76.DOI: 10.18280/mmc_a.910207.

[23] Lakshman Narayana Vejendla , A Peda Gopi and N.Ashok Kumar,(2018)," Different techniques for hiding the text information using text steganography techniques: A survey", Ingénierie des Systèmes d'Information, Vol.23, Issue.6,pp.115-125. DOI: 10.3166/ISI.23.6.115-125.

[24] A Peda Gopi and Lakshman Narayana Vejendla (2018), "Dynamic load balancing for client server assignment in distributed system using genetic algorithm", Ingénierie des Systèmes d'Information, Vol.23, Issue.6, pp.87-98.DOI: 10.3166/ISI.23.6.87-98.

[25] Lakshman Narayana Vejendla and Bharathi C R,(2017),"Using customized Active Resource Routing and Tenable Association using Licentious Method Algorithm for secured mobile ad hoc network Management", Advances in Modeling and Analysis B, Vol.60, Issue.1, pp.270-282. DOI: 10.18280/ama_b.600117.

[26] Lakshman Narayana Vejendla and Bharathi C R,(2017),"Identity Based Cryptography for Mobile ad hoc Networks", Journal of Theoretical and Applied Information Technology, Vol.95, Issue.5, pp.1173-1181. EID: 2-s2.0-85015373447.

[27] Lakshman Narayana Vejendla and A Peda Gopi, (2017)," Visual cryptography for gray scale images with enhanced security mechanisms", Traitement du Signal,Vol.35, No.3-4,pp.197-208. DOI: 10.3166/ts.34.197-208.

[28] A Peda Gopi and Lakshman Narayana Vejendla, (2017)," Protected strength approach for image steganography", Traitement du Signal, Vol.35, No.3-4,pp.175-181. DOI: 10.3166/TS.34.175-181.

[29] Lakshman Narayana Vejendla and Bharathi C R,(2016),"Secured Key Production and Circulation in Mobile Ad hoc Networks Using Identity Based Cryptography", International conference on Engineering and Technology, Vol.1, pp.202-206.

[30] A Peda Gopi and Lakshman Narayana Vejendla, (2019)," Certified Node Frequency in Social Network Using Parallel Diffusion Methods", Ingénierie des Systèmes d' Information, Vol. 24,No. 1, 2019,pp.113-117..DOI: 10.18280/isi.240117.

[31] Lakshman Narayana Vejendla and A Peda Gopi, (2019)," Avoiding Interoperability and Delay in Healthcare Monitoring System Using Block Chain Technology", Revue d'Intelligence Artificielle , Vol. 33,No. 1, 2019,pp.45-48.

[32] Boneh, Dan, and Matthew Franklin. "Identity-based encryption from the Weil pairing." SIAM journal on computing 32.3 (2003): 586-615.

[33] Hu, Hao, et al. "REPLACE: A reliable trust-based platoon service recommendation scheme in VANET." IEEE Transactions on Vehicular Technology 66.2 (2016): 1786-1797.

A Multi-Objectives Optimization to Develop the Mobile Dimension in a Small Private Online Course (SPOC)

Naima BELARBI[1]
Abdelwahed NAMIR[2]
Laboratory of Technological Information and Modelisation (LTIM)
Faculty of Sciences Ben M'Sik
University Hassan II
Casablanca
Morocco

Mohamed TALBI[3]
Observatory of Research in Didactics and University Pedagogy (ORDIPU)
Laboratory of Analytical Chemistry and Physical Chemistry of Materials
Faculty of Sciences Ben M'Sik
University Hassan II
Casablanca
Morocco

Nadia Chafiq[4]
Laboratory of Sciences, Information Communication and Education Technology (LAPSTICE)
Faculty of Sciences Ben M'Sik
University Hassan II
Casablanca
Morocco

***Abstract*—The impact of the mobile technology trend is being felt in several sectors today, including education. In this paper, we present an analysis of the development of the mobile dimension in a Massive Open Online Course (MOOC) or a Small Private Online Course (SPOC) as a decision-making problem among various approaches which cannot be ordered incontestably from the best to the worst. This is due to the fact that the various approaches to integrate the mobile dimension are different and that each solution presents both advantages and shortcomings from a technological point of view. The decision must be made on the basis of the end-users' requirements and usage. We propose to view this situation as a multi-objective optimization problem as the decision is a compromise between several conflicting objectives/criteria. The various approaches to the development of mobile access to a MOOC/SPOC are presented first and then compared using various criteria. Then we provide an analysis of the alternatives to find the non-dominant Pareto solutions.**

Keywords*—*Mobile dimension; MOOC/SPOC; multi-objective optimization; criteria; decision

I. Introduction

Nowadays, the impact of the mobile technology trend is felt in several sectors including education. In fact, the majority of users use their smartphones or laptops more for various uses. MOOCs and SPOCs are also affected by this challenge in order to meet users' expectations in terms of mobile access. MOOCs and SPOCs providers must also align their platforms with the requirements of mobile devices and mobile users. Indeed, having an online learning tool accessible to everyone, usable anytime and anywhere, and being able to easily access information and resources online is a very important aspect of mobile online learning which opens a new way to consider, explore and develop in MOOCs/SPOCs.

In the mobile world, the most popular approaches to mobile development are:

- Responsive Web Designing (RWD)
- Designing a mobile website (web app)
- Designing a mobile application
  - The design of a native mobile application
  - The design of a hybrid mobile application

These approaches are different in nature and the choice between them depends on the field of application and other considerations.

We propose to view the situation as a multi-objective optimization problem as the decision is the compromise between several conflicting objectives/criteria.

A Multi-objective optimization analysis is mainly at the origin of the field of Economic Balance and theories of well-being [1], but its application to several fields is constantly expanding thanks in particular to advances in data analysis techniques and algorithms. In fact, this optimization analysis was used in multitask learning [2], for solving engineering problems [3], in the metal forming industry [4] and to optimize laminated composite structures [5].

In this paper, we have first explored work concerning the fusion of both environments: MOOCs, SPOCS and mobile learning. We then listed the various mobile strategies of the main MOOC providers. We have also presented the various possibilities for the integration of the mobile dimension in the MOOC/SPOC environments, pointing out the advantages and the disadvantages of every approach. A comparison analysis between these different approaches based on the main criteria involved in decision making for a mobile solution has been presented.

Our main contribution is to demonstrate that the integration of the mobile dimension in a MOOC/SPOC is a multi-objective optimization problem. Indeed, it is not easy to find an "optimal" solution as the decision is a compromise among several conflicting objectives and this compromise is not perceived in the same way by all the decision-makers. This can therefore be considered as a multi-objective optimization problem. In the present work, we showed that the proposed four mobile solutions are an optimal set of non-dominant solutions since no improvements can be achieved for one objective without deteriorating the value of another one.

II. Related Works

A. MOOCs/SPOCs and Mobile Dimension

With the development of mobile technologies, MOOCs and SPOCs must meet this challenge in order to fulfill users' expectations in terms of mobile access. MOOC and SPOC providers also need to align their platforms to address the requirements of the mobile devices and users. The research that explored the combination of both learning formats, MOOCs and mobile learning, yielded some benefits, such as optimising the interaction and dialogue among learners, fostering collaborative learning, informal learning as well as lifelong learning [6] and the potential to extend the reach of MOOCs, improving access to learning materials and enabling new learning forms [7].

Choosing a strategy to integrate the mobile dimension into a MOOC/SPOC is not simple. Some research work provides some elements of comparison and feedback that can provide an idea of the different approaches to integrating the mobile dimension in a MOOC/SPOC.

In [8], the author reports that native or hybrid applications are no longer required to access a MOOC/SPOC that adheres to the principle of responsive design. However, native applications may offer functionality that cannot be provided in browser environments, including downloading videos for offline viewing. The authors also reported that a hybrid application, through openSAP, a hybrid application, has the advantage of combining native functionalities and web technologies in order to find a balance between maintainability, costs and user experience.

In [9], the authors argue that to create a mobile experience, there are always good reasons to adopt the two strategies, namely to create a native application and to create an optimized mobile website or a web application. They report that native apps are popular because they provide a native user experience, provide access to the hardware capabilities of mobile devices, and leverage the computing power of mobile devices. But the strategy of the native application requires the development of many applications as mobile platforms, which generates a tedious effort in terms of development times, diversified and specialized skill requirements. Regarding web apps, the authors report that despite the existence of tools that promise to reproduce the effects of native operation in the mobile browser, it is still not easy to obtain it. In addition, some features cannot be obtained through a mobile browser such as:

- Access to the hardware functions of mobile devices like integrated sensors, camera, etc.
- Access to mobile databases as SMS, address book, etc.
- The impossibility of running processes in the background in a mobile browser.

In this paper, we present an analysis of the problem of choosing the mobile solution to be implemented in a MOOC/SPOC as a decision-making problem among multiple solutions that can not be incontestably sorted from best to worst. We believe it is difficult to find an "optimal" solution as the decision involves a compromise between several conflicting objectives / criteria which are not perceived in the same manner by all decision makers. We suggest to consider this situation as a multi-objective optimization problem [10].

B. MOOC Providers and the Mobile Access Dimension

Early MOOC platforms were designed for exclusive access via desktop computers and the size of the mobile access was not taken into account [11]. However, with the rise of mobile technologies and their ubiquity in society, it has become imperative for MOOC providers to rethink the design of their platforms and have a mobile strategy to meet the needs of different learners, especially those of the mobile generation. The majority of MOOC platform providers such as Edx, Coursera, Moodle, Udacity and Iversity, have started to integrate the mobile dimension by offering either platforms with a responsive design that adapts to different mobile browsers and mobile device features, or dedicated native mobile applications that allow mobile access to their MOOC platforms [12]. According to [12], Coursera is ahead of the game in this area. Already in early 2014, it offered native applications addressing both iOS and Android platforms. Udacity also had two apps, for iOS and Android, but from 2019, both apps were removed from the App store and Google Play. This MOOC provider reports that it wants to focus more on improving the user experience on the desktop while offering a responsive design platform that can be adapted to all devices of different sizes. EdX has caught up with the mobile strategy and has proposed a mobile application starting in 2016 targeting the iOS platform and later Android.

In Table I, we present the main MOOC providers and their different mobile strategies.

TABLE. I. CONSIDERATION OF THE MOBILE DIMENSION BY THE MAIN MOOCS PLATFORM PROVIDERS

MOOC provider	Mobile web site	Responsive web design	Mobile app Yes/No	Supported mobile OS
Edx www.edx.org	No	Yes	Yes	iOS/Android
Udacity www.udacity.com	No	Yes	No	Available before 2019 on iOS and Android but from 2019, the 2 apps have been removed from App store and Google Play
Coursera www.coursera.org	No	Yes	Yes	iOS/Android
Udemy www.udemy.com	Non	Yes	Yes	iOS/Android
FutureLearn www.futurelearn.com	No	Yes	No	
Moodle www.moodle.org	No	Yes	Yes	iOS/Android
Khan Academy www.khanacademy.org	No	Yes	Yes	iOS/Android
Rwaq (Saoudi Arabia) www.rwaq.org	No	Yes	Yes	iOS/Android
uc@MOOC mooc.uca.ma (Cadi Ayyad University -Marrakech)	No	Yes	Yes	Android
Edraak JORDAN https://www.edraak.org	No	Yes	Yes	iOS/Android

III. THE DIFFERENT APPROACHES FOR THE MOBILE DEVELOPMENT

In the mobile world, the most popular approaches to mobile development are:

A. Responsive Web Designing (RWD)

In 2010, Ethan Marcotte introduced the concept of "Responsive Architecture" which means the ability of a physical space to adapt according to the presence of people passing through that space. This concept was later extended to the field of web design and currently refers to the ability of a website to automatically identify the display width of a page and to make the appropriate settings and adjustments necessary to guarantee a display adapted to each user's navigation device [13]. Thus, thanks to the techniques of the reactive Web, it is possible to create websites which adapt and optimize their content to the different navigation devices used by the user in order to improve their user experience whatever the terminal used. Indeed, these techniques allow:

- Adapting the layout to the different screen sizes for both large screen desktops and small phones;
- Automatically readjusting the content according to the width and / or height of the screen on which it is viewed;
- Resizing or changing the location of the images according to the screen resolution;
- Adjusting the structure and elements of the pages so that they can be viewed properly and without zooming on small screens such as those on smartphones or tablets;
- Hiding the elements which are not essential on screens having a small size;
- Making the links and buttons user-friendly for mobile users;
- And respecting users' preferences and personal parameters such as geolocation, device orientation, etc.

Developing a MOOCs / SPOCs platform with a responsive design has the advantage first of all by reducing the cost of development since it is a question of having a single website and offering the same content. Therefore, whatever the device used to follow the MOOCS/SPOCs, the user will have the same experience and follow the same content. In this case, it is not required to master several learning environments of the same MOOC/SPOC. This option also makes maintenance easier and avoids the risk of duplicate and different content.

On the other hand, the referencing of the platform in the search engine results page (SERP) of search engines is kept since it is the same and unique URL of the platform.

A platform of MOOCs/SPOCs in responsive design nevertheless has some drawbacks. We mainly cite the content download time which is longer on mobile devices. Indeed, with Responsive Design, all the elements making up a page are loaded, which has the consequence of reducing the performance of the MOOC/SPOC site on mobile devices.

Responsive Design is an interesting One Web approach if we want the MOOC/SPOC platform to be present on all terminals: Desktop, tablet, television, Smartphone, etc. and especially allows distributing a single and same content on all of these interfaces by reducing development costs while improving the user experience on mobile devices.

B. The Mobile Website

Mobile websites are another alternative for offering the content of an ergonomic MOOC/SPOC suitable for mobile devices with small screens. It is an approach which consists in developing a mobile site independent of the classic MOOC/SPOC website and which offers its own content, its own functionalities and its own organization. It has been developed so that its ergonomics are simple and suitable for mobile phones with short and targeted content for users. Touch navigation should be optimized through the use of icons to make it easier to use on mobile.

- A mobile site is accessible from the web browser, by entering the site's URL in the navigation bar, hence the need for an internet connection to be able to access it.

- There are different approaches for the development of a mobile website which can be considered as the light version of a standard website:
  - The "mobile first" approach, which consists of designing the mobile version of a site as a priority and then gradually adapting it to larger screens by adding effects and functionalities.
  - The classic approach, which is contrary to the concept of "mobile first" because it consists of first developing a standard website and then gradually adapting it to accommodate display on smaller mobile screens by removing effects and functionality which risks degrading the website.

With the improvement of the user-friendliness of mobile browsers, notably Mobile Safari on the iPhone since its first release in 2007 combined with the evolution of website design and development kits (HTML, JavaScript, CSS), display web pages and the readability of their content on small mobile screens has been remarkably improved, especially with the introduction of particular mobile functions such as touch actions, zoom, icons, etc. [14]. Consequently, "web applications", or "web apps" are sites created exclusively for mobiles and use the mobile browser as a runtime environment, thereby taking advantage of the browser's good support for mobile platforms. The "web apps" provide interactivity and navigation adapted for mobiles and an operation which is intended to be close to native mobile applications. The strategy of developing a mobile site enables the deployment of the content of a MOOC/SPOC on the various mobile platforms equipped with an internet browser. This has the advantage of offering redesigned content in order to adapt it to mobile platforms and more suitable functionalities for mobile interfaces. Web applications do not require installation or subsequent upgrades; updates are transparent to the end user. However, this approach also has drawbacks or constraints, in particular: the need for additional work to rethink the content to be deployed on mobile:

- The need for additional content development time to adapt it to mobile,
- The consequent double maintenance of two different environments of the same MOOC/SPOC,
- The need for the user to master two different learning environments depending on the terminal he uses to access the same MOOC/SPOC.
- Another problem is the additional time required to render web pages on mobile and the additional cost required to download the web page from the Internet.
- The need for a permanent internet connection to allow the user to follow a MOOC/SPOC on mobile [15].
- On the other hand, Web applications do not offer full access to the native functionality of smartphones and allow only limited access to the hardware and data of the mobile device; even if today, many software libraries promise to develop Web applications simulating the functionalities of native applications, such as JQuery Mobile, Sencha Touch, etc. In addition, the HTML5 standard promises access to the hardware and software components of the device through a number of APIs (Application Programming Interface).

C. *Mobile Applications (Mobile App)*

A mobile application or "mobile app" is a full-fledged program that installs and runs through the operating system of a mobile phone. It can be installed using an installation file directly on mobile or downloaded by the user via an online store such as App store, Google Play, Windows Phone Store or BlackBerry App World for free or for a fee. Once installed, a mobile app is permanently present on the mobile device and its execution does not generally require an internet connection, apart from certain mobile applications which are developed in connected mode.

The development of mobile applications has evolved with the development of the mobile device market since 2007, notably Apple's famous smartphones and iPhones, for which the first mobile applications were enormously successful both technically and commercially. But from 2010, mobile applications for the Android system began to experience a strong evolution [16]. For example, at the start of 2017, the App Store had around 2.2 million applications compared to 3.6 million applications listed in Google Play.

A mobile application is adapted to mobile screens and offers functionalities, ergonomic and graphic rendering that can be very advanced. Mobile applications offer simple use and navigation that is very suitable for mobile terminals characterized mainly by their small size.

Mobile apps are also marking another revolution with the mobile Internet. Indeed, the mobile Internet allows a mobile application to be embedded in the mobile, that is to say that the content is stored locally on the mobile. They can thus be instantly accessible or refreshed dynamically via network and server connections if the application is running in "connected" mode.

1) Types of mobile application: Mobile development is still faced with the dilemma of choosing between two alternatives, namely:

a) Native mobile applications that target a specific mobile platform such as IOS, Android, Microsoft, Windows Mobile, Symbian, BlackBerry, etc.

b) And hybrid applications which consist of developing a single code (Once code Principle) in standard web technologies (HTML5, JavaScript and CSS) and exploiting it on various existing mobile platforms.

2) Native development: The development of a native application means a specific development for each mobile OS (iOS, Android, Windows phone, etc.). This approach thus strongly depends on the target platform and imposes particular constraints such as the use of Integrated Development Environments (IDE), API (Application Programming Interface) and specific programming languages for each mobile platform. In Table II, we cite the different skills required in terms of development languages to be used according to the targeted mobile platforms.

TABLE. II. DEVELOPMENT SKILLS CORRESPONDING TO DIFFERENT MOBILE PLATFORMS

Mobile operating system	Required languages
Apple IOS	Objective C, Swift
Google Android	Java, kotlin
Symbian	C++, Python, HTML/CSS/JS
Windows mobile, Windows 7 Phone	C++, C#
Blackberry (ex RIM)	Java
Samsung bada	C++

3) Hybrid development: A hybrid application is a mobile application that operates halfway between a native application and a web app. Like native applications, it downloads from a store, launches like native applications and does not require an internet connection. It runs like a web application in a browser with the difference that the browser is part of the hybrid application, unlike a web app whose source code is always downloaded from the web.

4) Advantages and limitations: Hybrid applications are primarily designed using web technologies such as HTML5, JavaScript and CSS. Since 2015, the market for hybrid applications has grown considerably, especially with the evolution of mobile development frameworks that provide impressive results in terms of design, functionality and user experience very close to that obtained with native development.

The native application development process is the right way to deploy mobile-ready applications that deliver a rich user experience, fast performance, consistent look and full access to mobile platform hardware and data. Indeed, the unique technology environment of native applications enables the exploitation of native mobile functionality and takes full advantage of the specificity of mobile devices. For example, a native:

- Allows a better integration with the mobile system interface (pixel-accurate positions of the action buttons at the bottom of the screen, virtual keyboard, etc.);
- Offers the ability to use the mobile's hardware devices (GPS, camera, etc.);
- Provides interaction with "contacts, calendar, SMS/MMS functions" etc.
- Makes it possible to exploit alert and notification functions that are particularly symbolic of the advanced capabilities of mobile applications,
- And permits to take into account local user settings such as default language, portrait or landscape display mode offering the most personal user experience on the mobile.

However, there are disadvantages to this approach. First of all, it requires advanced technical skills and inevitably leads to an increase in development time because it is impossible to reuse the source code produced for one platform on another, since each system requires its own code. This leads to an increase in maintenance costs, which represents a real challenge for the native development approach.

As far as hybrid development is concerned, this approach allows single code writing and does not require any detailed knowledge of the target platform so that allows a multi-platform deployment. This makes it a very attractive model for several reasons, including the advantage of optimizing development time and effort, simplifying maintenance and deployment processes [17].

Despite the obvious advantages of hybrid mobile applications, there are also some disadvantages, such as limited access to the hardware features of mobile devices and reduced performance, especially in terms of speed and user experience as with native applications [17].

IV. COMPARATIVE ANALYSIS BETWEEN THE DIFFERENT APPROACHES TO MOBILE INTEGRATION IN A MOOC/SPOC

In Table III, we present a summary of the different strategies for integrating the mobile dimension in a MOOC/SPOC and some elements of comparison and decision-making criteria.

V. A MULTI-OBJECTIVE OPTIMIZATION ANALYSIS TO CHOOSE THE MOBILE SOLUTION

To set up a mobile strategy in a MOOC/SPOC, it appears that wanting to increase the performance of the mobile solution increases development time (Fig. 1). We are therefore faced with a dilemma where we are unable to improve the Performance criterion without deteriorating the development time criterion. This is typical of multi-objective optimization problems that consist in looking for the best compromise between the different objectives to be optimized [10]. For our analysis, we have chosen a rating scale from 1 to 5 to indicate the smallest to the greatest degree (Table IV).

Formally, a multi-objective optimization problem is a problem described as follows:

Minimize $F(x) = (f_1(x), f_2(x), \ldots\ldots, f_m(x))$ for $x \in X$ and knowing that $g_k(x) \le 0$. Where:

- $x = (x_1, x_2, \ldots, x_n)$ is the vector of n feasible solutions x_i with $1 \le i \le n$
- $x \in X$ the set of solutions
- $F(x)$ is the objective vector, where m is the number of objective functions with $m \ge 2$
- f_j is an objective to be optimized (minimized or maximized) where $1 \le j \le m$ knowing that maximizing f_j means minimizing $-f_j$.
- g_k represents a constraint that limits the values that solutions can take, $g_k(x) \le 0$ where: $1 \le k \le p$.

The solution of a multi-objective problem is not unique, but there is a set of solutions called "non-dominated set". These solutions are those where improvement in one objective systematically leads to deterioration in the performance of at least one other objective.

TABLE. III. COMPARISON OF THE DIFFERENT APPROACHES FOR THE DEVELOPMENT OF THE MOBILE DIMENSION IN A MOOC/SPOC

Criterion	**Native mobile app**	**Hybrid mobile app**	**Web app**	**Responsive design website**
Development	Code specific to each target platform	Unique code	-Technologies of the Web (HTML / Javascript / CSS)	LMS + responsive design theme
Skills required for mobile development	Highly advanced	advanced	Moderately advanced	Some settings at the LMS level -
Development time	Very long (Double)	Long (double)	Medium to long (Double)	Medium
Compatibility on platforms	Compatible on a single platform	Compatible on several platforms	online	online
Deployment	On a store	On a store	On Internet	On Internet
Need for Installation	Yes	Yes	No	No
Offline availability	Yes	Yes	No	No
Access to mobile equipment and data	Yes	Yes	Yes but limited	No
UX user experience	Yes complete	Yes but limited	Simulated Can be Enhanced by APIs /libraries Same ergonomics for all platforms	Web responsive
Performance	Native / custom	Simulated / Limited	Reduced Depends on the quality of the network, Depends on the browser Requires more Code and Resource Download Time	Low (time to download resources (videos, pictures, etc.)).
Distribution	On a store. Requires a validation at each update	On a store. Requires a validation at each update	On the web	On the web
Updates	requires reinstallation	requires reinstallation	Transparent	Transparent
Maintenance	Double	Double	Double	One site to maintain
MOOC/SPOC referencing	A mobile application + A site for the MOOC/SPOC	A mobile application + A site for the MOOC/SPOC	Two URLs: one for the mobile site and one for the MOOC/SPOC site.	A unique URL for the MOOC/SPOC site.
Price to market	Hight+	Hight+	Hight	Medium

TABLE. IV. RATING SCALE

Rating scale	
5	Very good
4	Well
3	Medium
2	Bad
1	Very bad

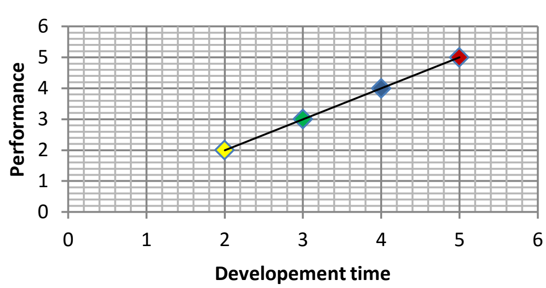


Fig. 1. Mobile Solutions: Performance vs. Development Time.

In the next section, we formulate the problem of choosing the mobile solution to be implemented in a MOOC/SPOC as an optimization problem with several objectives.

VI. DESCRIPTION OF THE PROBLEM OF CHOOSING THE MOBILE STRATEGY IN A SPOC

The problem of choosing the mobile strategy in a MOOC/SPOC can be described as follows:

- X A set of the mobile solutions $X = (x_1, x_2, x_3, x_4)$
- x_1: Native Mobile app (NM);
- x_2: Hybrid Mobile app (HM);
- x_3: Mobile Website (MW)
- x_4: Responsive Website (RW).
- The Several objectives to minimize or maximize: $f_j(x)$ where: $1 \leq j \leq 6$
- $f_1(x)$: Development time to minimize (DT)
- $f_2(x)$: Performances to maximize (PR) (which is equivalent to minimize $-f_2(x)$

- f_3 (x): User experience to maximize (UX)
- f_4 (x): Maintenance to be minimized (MN)
- f_5 (x): Update Transparent to the user to maximize (or minimize according to the designer's vision) (UT)
- f_6 (x): Offline access to maximize or minimize depending on the designer's vision (OF)

For illustrative purposes, Table V provides an example of the costs of the different criteria selected for each mobile solution on a scale of 1 to 5.

A. Illustration for Two Objectives

We will find the Pareto dominance relations [18] of the solutions x_i for the case of the following two objectives:

- The Development Time f_1 (to be minimized)
- The performance f_2 (to be maximized) which is equivalent to minimizing $-f_2$. (Inverse notation from 5 to 1: from the weakest to the strongest)

Let's call that a feasible solution $x \in X$ is "optimal Pareto" [19] or "not dominated" if and only if there is no solution $y \in X$ such that y dominates x.

It is said that a solution y "dominates" a solution z in the case of goal minimization, if and only if:

$\forall\ i \in [1..m],\ f_i(y) \leq f_i(z)$ and $\exists\ j \in [1..\ m]$ such as

$f_j(y) < f_j(z)$.

Thus, any solution in the Pareto set can be considered optimal since no improvement can be made on one objective without degrading the relative value of another objective. These solutions form what is called "the Pareto front".

In Fig. 2, we represent the case of the two-objective problem taken for illustration. We can notice that the four moving alternatives form the "Pareto front" since they are all non-dominated solutions.

TABLE. V. OBJECTIVES TO MINIMIZE

Mobile Solution/Objective	**f_1**	**$-f_2$**	**$-f_3$**	**f_4**	**$-f_5$**	**$-f_6$**
NM	5	1	1	5	5	1
HM	4	2	2	3	5	1
MW	3	3	3	3	1	5
RW	2	4	4	1	1	5

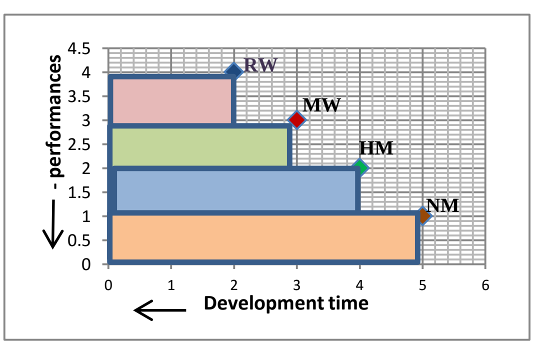


Fig. 2. Non Dominated Mobile Solutions.

B. Multi-Objective Case

In Fig. 3, we have represented the mobile solutions according to their ratings in relation to the different objectives [1] we have considered.

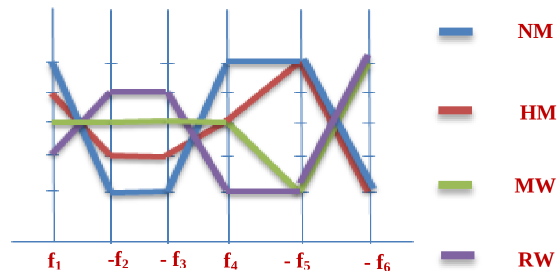


Fig. 3. Case of Several Objectives to be Minimized.

By comparing the solutions two by two, we can note that no one solution dominates the other:

- NM, HM: The native mobile solution minimizes $-f_2$ but maximizes f_1
- HM, NM: Mobile hybrid solution minimizes f_4 but maximizes $-f_3$
- NM, MW: Native mobile solution minimizes $-f_2$ but maximizes f_1
- MW, NM: Mobile web solution minimizes f_1 but maximizes $-f_2$
- NM, RW: Native mobile solution minimizes $-f_2$ but maximizes f_1.
- RW, NM: Responsive web solution minimizes f_4 but maximizes $-f_3$.
- HM, MW: Mobile hybrid solution minimizes $-f_2$ but maximizes $-f_5$.
- MW, HM: Mobile web solution minimizes f_1 but maximizes $-f_3$.
- HM, RW: Mobile hybrid solution minimizes $-f_2$ but maximizes f_4.
- RW, HM: Responsive site solution minimizes f_1 but maximizes $-f_3$.
- MW, RW: Mobile site solution minimizes $-f_2$ but maximizes f_4.
- RW, MW: Responsive site solution minimizes f_4 but maximizes $-f_3$.

Therefore, the four mobile solutions form an optimal set of solutions since no improvement can be made on one objective without degrading the relative value of another objective.

VII. CONCLUSION

Today, strengthening its presence on mobile is no longer an option but rather a necessity because the mobile has become an essential tool to access many services. MOOCs/SPOCs designers are also concerned with this challenge and must consider a solution to target all users, whatever the device used for access: desktop, laptop, smartphone, tablet.

The various approaches to integrate the mobile dimension in a MOOC/SPOC are different and each solution has advantages and disadvantages from a technical point of view. The choice must be made according to the needs and uses of the end users.

Based on the above analysis, we can admit with certainty, that the problem of choosing the mobile solution to be implemented in a SPOC is a problem of decision between several solutions that cannot be unquestionably ordered from the best to the worst. Moreover, it is difficult to find an "optimal" solution since the decision is a compromise between several conflicting objectives/criteria and this compromise is not seen in the same way by all decision-makers. We can thus consider this situation as a multi-objective optimization problem and in the present work. We found that the four mobile solutions proposed constitute an optimal set of non dominated solutions since no improvement can be made on one objective without degrading the relative value of another objective.

REFERENCES

[1] Sawaragi, Y., Nakayama, H., & Tanino, T. (Eds.). (1985). Theory of multiobjective optimization. Elsevier.

[2] Sener, O., & Koltun, V. (2018). Multi-task learning as multi-objective optimization. Advances in Neural Information Processing Systems (pp. 527-538).

[3] Mirjalili, S., Jangir, P., & Saremi, S. (2017). Multi-objective ant lion optimizer: a multi-objective optimization algorithm for solving engineering problems. Applied Intelligence, 46(1), 79-95.

[4] Ejday, M. (2011). Optimisation Multi-Objectifs à base de Métamodèle pour les Procédés de Mise en Forme (Doctoral dissertation).

[5] Irisarri, F. X., Bassir, D. H., Carrere, N., & Maire, J. F. (2009). Multiobjective stacking sequence optimization for laminated composite structures. Composites Science and Technology, 69(7-8), 983-990.

[6] De Waard, I., Koutropoulos, A., Hogue, R. J., Abajian, S. C., Keskin, N. Ö., Rodriguez, C. O., & Gallagher, M. S. (2012). Merging MOOC and mLearning for Increased Learner Interactions. International Journal of Mobile and Blended Learning, 4(4), 34–46. doi:10.4018/ jmbl.2012100103.

[7] Sharples, M., Delgado Kloos, C., Dimitriadis, Y., Garlatti, S., & Specht, M. (2015). Mobile and accessible learning for MOOCs. Journal of interactive media in education, 1(4), 1-8.

[8] Renz, J., Staubitz, T., & Meinel, C. (2014). MOOC to Go. International Association for the Development of the Information Society.

[9] Harb, E., Kapellari, P., Luong, S., & Spot, N. (2011). Responsive web design. Version 6.

[10] Gagné, C., Gravel, M., & Price, W. L. (2004). Optimisation multi-objectifs a l'aide d'un algorithme de colonie de fourmis. INFOR: Information Systems and Operational Research, 42(1), 23-42.

[11] Dalipi, F., Imran, A. S., Idrizi, F., & Aliu, H. (2016). An Analysis of Learner Experience with MOOCs in Mobile and Desktop Learning Environment. Advances in Intelligent Systems and Computing, 393–402. doi:10.1007/978-3-319-42070-7_36.

[12] Taneja, S., & Goel, A. (2014). MOOC providers and their strategies. International Journal of Computer Science and Mobile Computing, 3(5), 222-228.

[13] Peng, W., & Zhou, Y. (2015, July). The Design and Research of Responsive Web Supporting Mobile Learning Devices. In 2015 International Symposium on Educational Technology (ISET) (pp. 163-167). IEEE.

[14] Godwin-Jones, R. (2011). Mobile apps for language learning. Language Learning & Technology, 15(2), 2-11.

[15] Xanthopoulos, S., & Xinogalos, S. (2013, September). A comparative analysis of cross-platform development approaches for mobile applications. In Proceedings of the 6th Balkan Conference in Informatics (pp. 213-220). ACM.

[16] Dekhane, S., & Tsoi, M. (2012). Designing a Mobile Application for Conceptual Understanding: Integrating Learning Theory with Organic Chemistry Learning Needs. International Journal of Mobile and Blended Learning, 4(3), 34–52. doi:10.4018/ jmbl.2012070103.

[17] Malavolta, I., Ruberto, S., Soru, T., & Terragni, V. (2015, May). Hybrid mobile apps in the google play store: An exploratory investigation. In Proceedings of the second acm international conference on mobile software engineering and systems (pp. 56-59). IEEE Press.

[18] Marler, R. T., & Arora, J. S. (2004). Survey of multi-objective optimization methods for engineering. Structural and multidisciplinary optimization, 26(6), 369-395.

[19] Pareto, V. (1964). Cours d'économie politique (Vol. 1). Librairie Droz.

A Fuzzy Multi-Objective Covering-based Security Quantification Model for Mitigating Risk of Web based Medical Image Processing System

Abdullah Algarni[1], Abdulaziz Attaallah[3]
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah
Saudi Arabia

Masood Ahmad[2], Alka Agrawal[4]
Rajeev Kumar[5,*], Raees Ahmad Khan[6]
Department of Information Technology
Babasaheb Bhimrao Ambedkar University
Lucknow, Uttar Pradesh, India

***Abstract*—Medical image processing is one of the most active research areas and has big impact on the health sector. With the arrival of intelligent processes, web based medical image processing has become simple and errorless. Web based application is now used extensively for medical image processing. Large amount of medical data is generated daily with more and more data being shared over public and private networks for the diagnosis of diseases through the web based image processing systems. Medical images like that of the CT (Computed Tomography) scan, MRI (Magnetic Resonance Imaging), X-Ray and Ultrasound images, etc., contain highly personal data of the patients. This data needs to be secured from intruders. Medical images are more sensitive to external interruption and manipulation in data may cause changes in the result. Data breaches in medical cases can lead to wrong diagnosis or even more fatal possibilities with life threatening results. So, security in web based medical image processing is a major issue. However, ensuring security for the medical images while preserving the characteristics of confidentiality, integrity, availability, etc., of medical images poses a major challenge. Working towards a feasible solution, in this study, authors are using a list of criteria for checking security level of the web based image processing system. We propose Fuzzy Analytic Hierarchy Process (FAHP) combined with Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) in the list of criteria that affect the security assessment in medical image processing. At the results we see that FAHP-TOPSIS produce good results in security checking in web based medical image processing system. At the data analysis section all the steps showed which is involved in our model.**

Keywords*—*Web based medical image processing; fuzzy analytical hierarchy process; TOPSIS method; security management

I. Introduction

Medical Image Processing is the most critical aspect of the diagnosis of a disease. The major benefit of the medical image processing is that it facilitates in detection of the disease in its early stages [1]. If the image processing is not done properly, it may lead to wrong diagnosis of the disease. Before the diagnosis, hospitals and doctors send the images or data for the image processing like storing, retrieving, denoising, etc. [2]. In this respect, the hospitals generate huge amount of digital data [3]. A new technology called cloud computing has emerged as one of the potential solutions for processing the medical image data [2]. Most of the hospitals do not own data storage space because medical images are large in size, making data storage an impractical alternative. Hence, most of the hospitals or labs avail of the cloud computing technology. Cloud provides the space and security of the data for which the hospitals pay the requisite tariff[4].

While the recent advancement in technologies provide new means (like cloud based image processing, web based medical image processing etc.) to handle the medical images, they also compromise their security due to easy to retrieve, manipulation and replication [5][6]. Most of these technologies are highly vulnerable in the present cyber security context because of ineffective security mechanisms. When attackers find loopholes in the existing security technology, then it is imperative to design new security technology for making the systems more secure [7] [8].

Medical image security has gained significant space in the recent endeavours of cyber security experts, practitioners and academic researchers. Security involves the following aspects: Confidentiality (only authorize users can access patient data), Availability (data available at the time of Natural damage), integrity (show that medical data hasn't been changed) and authentication (information origin to be proven)[9]. Many methods can be applied to provide security for medical image like steganography, watermarking and encryption. Steganography and watermarking are used for authentication to prevent access to an unauthorized entity [6] [10]. There are a large number of image encryption methods because different applications require different levels of security [10] [11]. All of these securities techniques are used in transmission of data over the network [12]. Ample availability of digital data becomes an easy prey for the attackers to intrude upon, especially when the medical data is being sent over the network for processing the image, the attackers begin their attempts to trace the data [13]. Thus, the need for devising security mechanisms that afford optimum data integrity is becoming imminent by the day.

However before the use of security methods, security assessment is an even more important concern [14]. Most of the medical image systems are interlinked with Internet, and

*Corresponding Author

web applications [15]. These systems are easy targets of the attackers because limited mechanisms are applied to ensure system security and data privacy [9]. Most of the image processing devices like CT, X-RAY, MRI can also be invaded by the attackers. Medical devices also need security at the time of processing the images [11] [16]. In the present era most of the hospitals or diagnostics centers used web based application for storing the patient's data and processing on them [14] [5]. At the time of development of application, the developers do not focus on the security thereby unintentionally giving open access to the intruders [17][18].

In this context, for the assessment of cyber security risk, Pingchian Ma et al. proposed a hybrid model of AHP and FUZZY comprehensive Evaluation [16], this method is totally based on medical device security assessment. Limitation of Classical AHP is volatile scale of judgment and ranking problem in the device. In this paper author will present Fuzzy AHP- TOPSIS hybrid model to overcome the volatile scale of judgments and provide the ranks of the medical image processing system. Authors have done the comparative study with classical AHP and Fuzzy AHP and results displayed in the comparison section.

II. Literature Review

Most of the researchers are trying to find the security failure causes. In this literature review find out that security attributes shows an important role in medical image processing system security. CIA is the basic three poles of security which shows an important role in improving the security. Several factors like authentication, authorization, utility, possession and resilience remains which show an important role in medical image processing system security. Security related literatures are explained as: In 2019 A. Agrawal et.al. showing the sustainable security measurement on web application using. Agencies and development companies develop guidelines for making web applications design sustainable and secure.

In 2018 Yinghui et.al proposed a secure and privacy aware smart health system which is based on patient data privacy. Authors propose a secure resilient health system for protection and safe medical data transmission reduce risk if the private key becomes leakage. In 2018 Aqsa & Ricardo did a comprehensive study of security mapping in healthcare information systems. Authors find the issues in implementation and exploitation.

In2019 Shi & et.al develop a framework for privacy protection for health care big data management based risk access control. This is based on reliability of risk analysis of data in smart health care system patient data can be leak from three aspects: resource sensitivity, access behavior sensitivity and historical access. For risk assessment authors used fuzzy rule techniques which is used for decision making and providing guidance for improving healthcare system.

In 2018 Marwan et.al proposed a cloud based framework for securing medical image processing. Author used combined segmentation techniques and genetic algorithm for prevention of accidental disclosure of data. In 2018 Arun, Ashish & George provide a study on healthcare informatics and privacy. Author identify that not any standard body is available for identification genomic data as personal data. This method identifies and prevent from the access control.

M.Moayeri et.al(2015) done comparative study of Fuzzy AHP an TOPSIS methods for Math teachers selection. In this study authors shows that Fuzzy TOPSIS is better in comparison to classical AHP and TOPSIS.

In 2018Pingchuan Ma et al. proposed an FAHP model, for quantitative cyber security assessment, this model focusing on Medical imaging device security assessment the security in medical device and guidelines for manufacture and government to design secure devices. In 2018 M.Fatih & Gul used AHP-TOPSIS with Pythagorean Fuzzy Sets for security risk analysis. AHP Pythagorean Fuzzy used for expert judgment and TOPSIS Pythagorean Fuzzy for prioritization of identified risk. For risk analysis author used three parameters privacy, integrity and accessibility for risk analysis which shows that this methods improves effectiveness of classical risk analysis method. Alots of researcher work done in medical image secuirity but doesn't have work on medical image processing system security assessment. Ranking system in security of working system determine the longer security. In this paper author determine the ranking of security using FUZZY AHP-TOPSIS method discuss in next sections.

III. Security Assessment of Medical Image Processing System

Medical image processing system in modern medicine has become increasingly important; it is the best fit for a rapid and effective diagnosis. Actually, the medical image data provides useful information to doctors; assists in decision making and, as a result, improves treatment. Thus, any accidental change in medical image can negatively effect on the treatment [19]. Security of healthcare system/image processing system is essential part of the healthcare industry and the industry is intensifying its efforts in this direction.

All these findings of criteria satisfy the goal of assessment the security of medical image processing system. Security assessment criteria and goal show in the figure-1. For the security assessment, attributes of security assessment described below in detail.

- Confidentiality- Refers to minimal access and disclosure of data (or resource). A loss of confidentiality happens when the data is actually accessed by unauthorized user. Confidential information is not available to unauthorized access or for all users. Medical images also contain sensitive information which only the authorized person can access.

- Authentication- The main task of authentication is to validate the user, check the user's identity and match the data which is stored in the database. Authentication process provides access control for system by verifying, to see if user's given information matches with the database of the authorized users.

- Authorization- After the completion of identification and authentication process, the asset gives the permission to access the data. This process is called authorization.

- Availability- Data is said to be available if it is available in any circumstances when needed. In terms of medical image processing data should be available at over the traffic on website. The data should be manageable at any sustainable environment. Availability of data can be prevented due to power failure, hardware failure and software upgrades.

- Integrity- In terms of medical image processing, image data should not change. If the image data is changed at the processing time, then the results may be faulty. Data should not be altered in any circumstances and this is called data integrity. Data should be complete and unchanged from the original. Integrity maintains the accuracy of information.

- Utility-In the terms of medical image processing, image useful for sometime only. This is called utility. Old data is not useful for data utility purpose. Usefulness of data for a purpose is called utility of data. The difference between utility and availability is that the data is still available but no longer usable.

- Possession- Is to retaining the ownership and make data under control. If possession is lost then there would be control over the data. In terms of medical image processing, ownership of the data is imperative otherwise the original data would be lost or worst, be corrupted.

- Resilience- Resilience is the process of checking the resistance of medical image processing system to attacks. Resilience is implemented by using, OTP and encryption. Main task of resilience is to protect the entire system from attack, taking into observation all the unsafe components of the system.

Healthcare system design plays an important role in healthcare industry. While developing the healthcare system data, security should be the most prioritized concern. Through the security attributes we have designed the guidelines to develop the system. The rule set of medical image processing system security follows security attributes for assessment the security. Security assessment becomes difficult through the traditional method. To remove this difficulties and biasness, the FUZZY set theory used with AHP and TOPSIS give fruitful results.

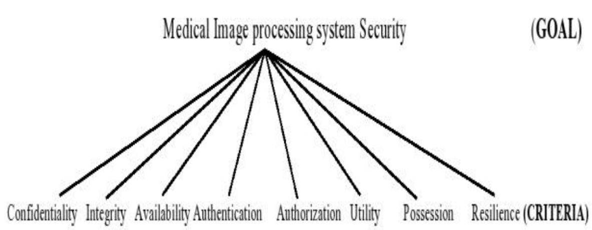


Fig. 1. Medical Image Processing System Security Attributes.

IV. Methodology

Different authors have analyzed medical image security. Changing the medical image processing security with Fuzzy AHP-TOPSIS is a new method for achieving both high security and user satisfaction. Besides, to achieve the goals as healthcare industry wants, we have used the Multi-Criteria Group Decision Making (MCGMD). In this section, we describe an approach for security assessment in medical Image processing system using Fuzzy AHP-TOPSIS. For the assessment of security, AHP approach is very suitable. But AHP faced the criticized due to unbalanced scale of judgments and it takes an exact value for decision making. To overcome these faults we have used Fuzzy AHP techniques for security assessment and Fuzzy TOPSIS used for providing the ranks of the systems. We took the sequence of steps of Fuzzy AHP-TOPSIS method to find the results which are shown as:

a) Fuzzy AHP: Fuzzy AHP is the approach used to calculate the weights of criteria; Fuzzy AHP represents problems in the hierarchy tree form with levels (goal and criteria). The top level shows the goal and objective. Second level shows the criteria and sub-criteria. The next step is building the Triangular Fuzzy Number (TFN) from the hierarchal structure. Triangular Fuzzy value is used for creating pair-wise comparison matrix.

Triangular Fuzzy membership value for pair wise comparison was employed by Chang[20]. In this paper, we adopted TFN, because they make calculation of membership functions easy and share out with fuzzy data. The TFN lies between 0 and 1. The linguistic values are divided as equally important, weakly important etc., and crisp values are shown as numeric 1,2,………..9., its membership function values are calculated by this equation (1-2):

$$\mu_a(x) = a \rightarrow [0,1] \tag{1}$$

$$\mu_a(x) = \begin{cases} \frac{x}{B-A} - \frac{l}{B-A} & x \in [A,B] \\ \frac{x}{B-C} - \frac{u}{B-C} & x \in [B,C] \\ 0 & Otherwise \end{cases} \tag{2}$$

Where, assigned A as lower, B as middle, and C as upper value equally in the triangular membership function. Figure 2 represent TFN value.

A TFN is shown in figure 2. Experts assigned the quantitative value to the linguistic terms in value; values are shown in table 1.

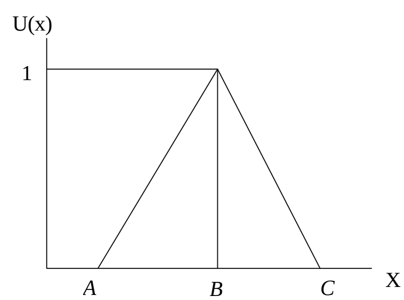


Fig. 2. Triangular Fuzzy Number.

TABLE. I. TFN SCALE

Security assessment Scale	Definitions	Membership function
1	Equally important	(1 ,1, 1)
3	Weakly important	(2 ,3, 4)
5	Fairly important	(4 ,5, 6)
7	Strongly important	(6 ,7, 8)
9	Absolutely important	(9 ,9, 9)
2 4 6 8	Intermediate values between two adjacent scales	(1 ,2, 3) (3 ,4, 5) (5 ,6, 7) (7 ,8, 9)

The equations (3 to 6) are used for converting the numerical data into TFN [19], Φ_{ij} is calculated depend on the geometric mean of specialists' observation for a particular similarity. TFN [Φ_{ij}] is calculated as:

$$\Phi_{ij} = (A_{ij}, B_{ij}, C_{ij}) \tag{3}$$

$$where\ A_{ij} \leq Bi_{ij} \leq C_{ij}$$

$$A_{ij} = min(J_{ijd}) \tag{4}$$

$$B_{ij} = (J_{ij1}, J_{ij2}, J_{ij3})^{\frac{1}{x}} \tag{5}$$

$$and\quad C_{ij} = max(J_{ijd}) \tag{6}$$

The geometric mean is calculated by multiplying and adding two fuzzy numbers. Equations (7-9) used to calculate geometric mean. Consider two TFNs P1 and P2, P1= (A_1, B_1, C_1) and P2= (A_2, B_2, C_2). Calculation of Geometric means shown as:

$$(A_1, B_1, C_1) + (A_2, B_2, C_2) = (A_1 + A_2, B_1 + B_2, C_2 + C_2) \tag{7}$$

$$(A_1, B_1, C_1) \times (A_2, B_2, C_2) = (A_1 \times A_2, B_1 \times B_2, C_1 \times C_2) \tag{8}$$

$$(A_1, B_1, C_1)^{-1} = (\frac{1}{C_1}, \frac{1}{B_1}, \frac{1}{A1}) \tag{9}$$

A pair-wise n x n comparison matrix is created by dividing the row element with column by this equation (10).

$$\widetilde{M^d} = [\tilde{N}^d_{11} \mathrm{N}^d_{12} .. \tilde{N}^d_{1n} \tilde{N}^d_{21} \tilde{N}^d_{22} \ldots . \tilde{N}^d_{2n} \cdots \tilde{N}^d_{n1} \tilde{N}^d_{n2} \ldots \tilde{N}^d_{nn}] \tag{10}$$

Where $\widetilde{N^d_{ij}}$ shows the d[th] experts give the importance of the i[th] fact over the j[th] fact. If more than one expert is present, then the average of each specialist is calculated by this equation (11).

In addition, we divide the Consistency Index by Random Index [(*RI*) *is generated from Saaty*] for calculating the Consistency Ratio (CR). This is shown in statement (11):

$$\tilde{N}_{ij} = \sum_{d=1}^{d} \quad \tilde{N}^d_{ij}/\mathrm{n} \qquad \text{n is the number of experts} \tag{11}$$

Next stage, Take the average of all factors in the hierarchy, here change the pair-wise comparison matrixes by this equation (12).

$$\tilde{M} = \lfloor \widetilde{\mathrm{N}_{11}} \ldots \widetilde{N_{1n}} \cdots \tilde{\mathrm{N}} \cdots \tilde{N}_{nn} \rfloor \tag{12}$$

After updating the pair-wise comparison matrix, with the help of equation (13) to calculate the fuzzy geometrical mean and fuzzy weights of every factor.

$$\tilde{O}_i = (\prod_{j=1}^{n} \quad \tilde{N}_{ij})^{\frac{1}{n}}, i = 1,2,3 \tag{13}$$

Next stage, we add all geometric mean values to find fuzzy weights by this equation (14).

$$\tilde{Q}_i = \tilde{O}_i \otimes (\tilde{O}_1 \oplus \tilde{O}_2 \oplus \tilde{O}_3 \ldots . \oplus \mathrm{O}_n)^{-1} \tag{14}$$

Next stage, we calculate the fuzzy average weight through equation (15).

$$P_i = \frac{\tilde{Q}_1 \oplus \tilde{Q}_2 \ldots . . \oplus \tilde{Q}_n}{n} \tag{15}$$

Further, we normalized the fuzzy weight through the equation (16).

$$Nw_i = \frac{P_i}{P_1 \oplus P_2 \oplus \ldots \ldots \oplus P_n} \tag{16}$$

After that, we can de-fuzzify the fuzzy weights to get crisp values; the de-fuzzification methods use the Centre of Area (COA) to calculate the BNP (Best Non-fuzzy Performance) value of the fuzzy weights by equation (17).

$$BNPwD1 = \frac{\lfloor (CQ1 - AQ1\ ) + (BQ1 - AQ1\ ) \rfloor}{3} + AQ1 \tag{17}$$

b) Fuzzy TOPSIS: TOPSIS is used in the scenario of performance value decision. It is not used in crisp value but instead in the linguistic value given by decision maker. We used linguistic terms like very poor, poor, fair, good and very good. Without the numerical value, it is tough to assign the rank. Instead of directly assigning the linguistic value, the decision maker used Fuzzy AHP for fuzzy values for weights for each criterion. In addition, Fuzzy AHP-TOPSIS approach is totally suitable for fixing group decision-making problems in fuzzy environments. Fuzzy AHP-TOPSIS technique is as follows:

- In the first step, Fuzzy AHP is using to calculating fuzzy choice weights by mathematical statement (1-16).
- At last, by this mathematical statement (18) and table 2 we design the fuzzy decision matrix.

$$\begin{array}{cc} & \begin{array}{ccc} C_1 & \ldots\ldots . & C_n \end{array} \\ \tilde{M} = \begin{array}{c} A_1 \\ \ldots \\ A_m \end{array} & \begin{bmatrix} \tilde{x}_{11} & \cdots & \tilde{x}_{1n} \\ \cdots & \ddots & \cdots \\ \tilde{x}_{m1} & \cdots & \tilde{x}_{mn} \end{bmatrix} \end{array} \tag{18}$$

Next stage, the standardized fuzzy decision matrix is represented by $\tilde{\mathrm{F}}$, simplified by the equation (19).

$$\tilde{F} = [\tilde{F}_{ij}]_{m \times n} \tag{19}$$

TABLE. II. LINGUISTIC SCALE FOR THE RATING

Linguistic Variable	*Corresponding Triangular Fuzzy Number*
Very poor (VP)	(0, 1, 3)
Poor (P)	(1, 3, 5)
Fair (F)	(3, 5, 7)
Good (G)	(5, 7, 9)
Very good (VG)	(7, 9,10)

After that, the standardization process can be done by the equation (20).

$$\tilde{F}_{ij} = \left(\frac{A_{ij}}{c_j^+}, \frac{B_{ij}}{c_j^+}, \frac{c_{ij}}{c_j^+}\right), \quad C_j^+ = max\{C_{ij}, i = 1,2,3..n\} \tag{20}$$

On the other hand, author opt for the supreme of C_j^+ is equal to 1 where and j = 1, 2, . . . , n; and worst case is equal to 0. The weighted fuzzy standardized decision matrix ($S\tilde{Q}$) is calculated by the equation (21).

$$\widetilde{SQ} = [s\tilde{q}_{ij}]_{m\times n} i = 1,2,..m; \; j = 1,2,3 \dots n \tag{21}$$

Where, $s\tilde{q}_{ij} = \tilde{F}_{ij} \otimes \tilde{Q}_{ij}$, thereafter, author represent the Fuzzy Positive-Ideal Solution (FPIS) and Fuzzy Negative-Ideal Solution (FNIS). Here components $s\tilde{q}_{ij}$ are standardized in positive TFN, shown by the standardized weighted fuzzy decision matrix shown in table 6. And positive TFN values lie between [0, 1].The FPIS R^+ (Supreme) and FNIS R^- (worst) are calculated by the equation (22-23).

$$R^+ = (\tilde{q}_1^*, \dots \tilde{q}_j^*, \dots \tilde{q}_n^*) \tag{22}$$

$$R^- = (\tilde{q}_1^*, \dots \tilde{q}_j^*, \dots \tilde{q}_n^*) \tag{23}$$

Where,

$$\tilde{q}_1^* = (1,1,1) \otimes \tilde{Q}_{ij} = (LQ_j, MQ_j, HQ_j) \text{ and } s\tilde{q}_{ij}^- = (0,0,0)$$

Through the FPIS and FNIS author evaluate the distance of every alternative. The area compensation technique is used for calculating the distances ($\tilde{D}_i^+ and \ \tilde{D}_i^-$) of each alternative from R^+ and R^- as shown by the equation (24-25).

$$\tilde{D}_i^+ = \sum_{j=1}^{n} \quad D(\widetilde{sq}_{ij}, s\tilde{q}_{ij}^*) i = 1,2,..m; \; j = 1,2,3 \dots n \tag{24}$$

$$\tilde{D}_i^- = \sum_{j=1}^{n} \quad D(s\tilde{q}_{ij}, s\tilde{q}_{ij}^*) i = 1,2,..m; \; j = 1,2,3 \dots n \tag{25}$$

Now, Closeness coefficients ($\widetilde{CoC}_i$) are calculated by the equation (26) in this stage, and the other option is developed to achieve the desire levels in each factor.

$$C\widetilde{oC}_i = \frac{\tilde{N}_i^-}{N_i^+ + \tilde{N}_i^-} = 1 - \frac{\tilde{N}_i^+}{\tilde{N}_i^+ + N_i^-}, i = 1,2, \dots, m \tag{26}$$

V. DATA ANALYSIS AND RESULTS

Mostly, Qualitative evaluation is good for evaluating the security of web based medical image processing system. It is typical to calculate web based image processing system security quantitatively. In recent years, Healthcare industries are trying to select high security medical image processing systems or devices [16]. In addition, medical image security plays an important role during processing (capture, store, retrieve, etc.) [5][10]. This research study proposes medical image security through Fuzzy AHP- TOPSIS approach[18]. The medical image security assessment attributes have been divided and explained by the authors in the previous parts of this paper shown in figure 1. We used the equations (1-26) for calculating the web based medical images processing security assessment using Fuzzy AHP-TOPSIS approaches, as was depicted through Table 1 by using equations (1 to 9). Thereafter, we performed the fuzzification where linguistic values were converted into numerical values and made the scale for linguistic terms. These values were used to construct a pair-wise comparison matrix. Equation (10) is used to calculate the pair-wise comparison matrix as shown in table 3.

In the next step, we calculate the fuzzy weights of factors with the help of the equations (11-13) shown in table(4).

In the present scenario of security assessment, it is important to analyze the effect of medical image processing security assessment in different options as per their requisite goals and criteria. Therefore, we are using 7 different web based projects pertaining to medical image processing system for hospitals in Varanasi, UP, India. Here, A1, A2, A3, and A4 show the hospital based image processing system. The remaining A5, A6, and A7 projects are web applications based diagnostic centers. Further, Hospital based project is represented by HB and diagnostic center based project represented by DC. Because of the security of the patient's data, all these web based medical projects are very sensitive. We are using table 2 for the technical data of the seven projects as shown in table 5. We calculated regularized fuzzy decision matrix by using the equations (18-20). We calculated weighted normalized fuzzy decision matrix as presented in table 6, by using the equation (21) and we calculated the closeness coefficient aspire level presented in table 7 based on the equation (22-26).

Overall satisfaction degree of project is classified in ranks; rank is obtained in between 1 to 7 as shown in Table 7. As observed, in this case study, rank order obtained in 7 alternatives finds as DS7 > HB2 > HB4 > DC6 > HB1 > DC5 > HB3.

TABLE. III. FUZZY AHP PAIR WISE MATRIX

	F1	F2	F3	F4	F5	F6	F7	F8	F9
F1	1.0000, 1.0000, 1.0000	1.0000, 1.0000, 1.1900	0.7200, 0.9000, 1.1500	0.5500, 0.7700, 1.0000	0.7700, 1.0000, 1.1500	0.7200, 0.9000, 1.1500	0.7700, 1.0000, 1.1500	0.5500, 0.7700, 1.0000	0.4500, 0.5900, 0.8500
F2	0.8700, 1.0000, 1.0000	1.0000, 1.0000, 1.0000	0.8500, 0.9000, 1.3500	0.5500, 0.7700, 1.0000	0.7700, 1.0000, 1.1500	0.7200, 0.9000, 1.1500	0.7700, 1.0000, 1.1500	0.5500, 0.7700, 1.0000	0.4700, 0.6200, 0.9000
F3	0.9000, 1.1500, 1.5000	0.7700, 1.1500, 1.3100	1.0000, 1.0000, 1.0000	0.7200, 0.9000, 1.0000	1.0000, 1.0000, 1.1500	1.0000, 1.0000, 1.0000	1.0000, 1.0000, 1.1500	0.7200, 0.9000, 1.0000	0.5900, 0.8500, 0.9000
F4	1.0000, 1.3500, 1.8500	1.0000, 1.3500, 1.8500	1.0000, 1.1500, 1.5000	1.0000, 1.0000, 1.0000	1.0000, 1.1500, 1.6500	1.0000, 1.1500, 1.5000	1.0000, 1.1500, 1.6500	1.0000, 1.0000, 1.0000	0.8000, 1.0000, 1.0000
F5	0.9000, 1.0000, 1.3500	0.9000, 1.0000, 1.3500	0.9000, 1.0000, 1.0000	0.6200, 0.9000, 1.0000	1.0000, 1.0000, 1.0000	0.9000, 1.0000, 1.0000	1.0000, 1.0000, 1.0000	0.6200, 0.9000, 1.0000	0.5300, 0.7400, 0.9000
F6	0.9000, 1.1500, 1.5000	0.9000, 1.1500, 1.5000	1.0000, 1.0000, 1.0000	0.7200, 0.9000, 1.0000	1.0000, 1.0000, 1.1500	1.0000, 1.0000, 1.0000	0.9000, 1.0000, 1.1500	0.7200, 0.9000, 1.0000	0.5900, 0.8500, 0.9000
F7	0.9000, 1.0000, 1.3500	0.9000, 1.0000, 1.3500	0.9000, 1.0000, 1.0000	0.6200, 0.9000, 1.0000	1.0000, 1.0000, 1.0000	0.9000, 1.0000, 1.1500	1.0000, 1.0000, 1.0000	0.6200, 0.9000, 1.0000	0.5100, 0.7100, 0.8500
F8	1.0000, 1.3500, 1.8500	1.0000, 1.3500, 1.8500	1.0000, 1.1500, 1.5000	1.0000, 1.0000, 1.0000	1.0000, 1.1500, 1.6500	1.0000, 1.1500, 1.5000	1.0000, 1.1500, 1.6500	1.0000, 1.0000, 1.0000	0.8000, 1.0000, 1.0000
F9	1.3100, 1.8100, 2.3100	1.1500, 1.6500, 2.1500	1.1500, 1.3100, 1.8100	1.0000, 1.0000, 1.3100	1.1500, 1.4600, 1.9600	1.1500, 1.3100, 1.8100	1.3100, 1.6200, 2.1200	1.0000, 1.0000, 1.3100	1.0000, 1.0000, 1.0000

TABLE. IV. WEIGHTS OF FACTORS

Factors	Weights	BNP	Rank
F1	0.0700, 0.1000, 0.1400	0.0840	6
F2	0.0700, 0.1000, 0.1400	0.0810	7
F3	0.8000, 0.1100, 0.1400	0.0950	5
F4	0.0900, 0.1200, 0.1800	0.1500	2
F5	0.0700, 0.1000, 0.1400	0.0800	8
F6	0.0800, 0.1100, 0.1400	0.1100	4
F7	0.0700, 0.1000, 0.1400	0.0800	9
F8	0.0900, 0.1200, 0.1800	0.1500	3
F9	0.1000, 0.1500, 0.2200	0.1700	1

TABLE. V. SUBJECTIVE COGNITION RESULTS

Alternatives/ Factors	F1	F2	F3	F4	F5	F6	F7	F8	F9
HB1	6.3800, 8.3800, 9.6900	5.0000, 7.0000, 9.0000	7.0000, 9.0000, 10.0000	7.6200, 9.3100, 10.0000	6.2400, 8.2400, 9.6200	6.2400, 8.2400, 9.6200	4.3800, 6.3800, 8.3800	5.6200, 7.6200, 9.3100	5.6200, 7.6200, 9.3100
HB2	5.0000, 7.0000, 9.0000	3.7600, 5.7600, 7.7600	4.2400, 6.2400, 8.2400	5.0000, 7.0000, 9.0000	5.6200, 7.6200, 9.3100	3.6200, 5.6200, 7.6200	4.2400, 6.2400, 8.2400	7.0000, 9.0000, 10.0000	5.6200, 7.6200, 9.3100
HB3	7.6200, 9.3100, 10.0000	5.6200, 7.6200, 9.3100	9.0000, 10.0000, 10.0000	5.6200, 7.6200, 9.3100	7.6200, 9.3100, 10.000	0.0000, 0.6200, 2.2400	5.0000, 7.0000, 9.0000	0.3100, 1.6200, 3.6200	0.6200, 2.2400, 4.2400
HB4	7.0000, 9.0000, 10.0000	5.7600, 7.7600, 9.3800	3.0000, 5.0000, 7.0000	1.7600, 3.7600, 5.7600	1.6200, 3.6200, 5.6200	3.0000, 5.0000, 7.0000	5.6200, 7.6200, 9.3100	3.0000, 5.0000, 7.0000	5.6200, 7.6200, 9.3100
DC5	3.0000, 5.0000, 7.0000	5.6200, 7.6200, 9.3100	3.0000, 5.0000, 7.0000	7.0000, 9.0000, 10.0000	6.3800, 8.3800, 9.6900	3.0000, 5.0000, 7.0000	5.0000, 7.0000, 9.0000	0.3100, 1.6200, 3.6200	3.6200, 5.6200, 7.6200
DC6	3.0000, 5.0000, 7.0000	5.0000, 7.0000, 9.0000	5.6200, 7.6200, 9.3100	6.2400, 8.2400, 9.6200	3.0000, 5.0000, 7.0000	3.0000, 5.0000, 7.0000	3.6200, 5.6200, 7.6200	7.0000, 9.0000, 10.0000	3.0000, 5.0000, 7.0000
DC7	5.0000, 7.0000, 9.0000	5.0000, 7.0000, 9.0000	3.6200, 5.6200, 7.6200	3.6200, 5.6200, 7.6200	3.7600, 5.7600, 7.7600	3.6200, 5.6200, 7.6200	7.0000, 9.0000, 10.0000	7.0000, 9.0000, 10.0000	5.7600, 7.7600, 9.3800

TABLE. VI. WEIGHTED STANDARDIZE FUZZY DECISION

Alternative/ Factors	F1	F2	F3	F4	F5	F6	F7	F8	F9
HB1	0.00900, 0.01100, 0.01300	0.00700, 0.01000, 0.01300	0.01300, 0.01700, 0.01900	0.01200, 0.01500, 0.01600	0.00800, 0.01100, 0.01300	0.00400, 0.00600, 0.00700	0.00300, 0.00500, 0.00600	0.00500, 0.00700, 0.00800	0.00800, 0.01100, 0.01300
HB2	0.00700, 0.01000, 0.01200	0.00500, 0.00800, 0.01100	0.00800, 0.01200, 0.01500	0.00800, 0.01100, 0.01400	0.00700, 0.01000, 0.01200	0.00300, 0.00400, 0.00500	0.00300, 0.00500, 0.00600	0.00600, 0.00800, 0.00900	0.00800, 0.01100, 0.01300
HB3	0.01000, 0.01300, 0.01400	0.00800, 0.01100, 0.01400	0.01700, 0.01900, 0.01900	0.00900, 0.01200, 0.01500	0.01000, 0.01200, 0.01300	0.00000, 0.00000, 0.00200	0.00400, 0.00500, 0.00700	0.00000, 0.00100, 0.00300	0.00100, 0.00300, 0.00600
HB4	0.01000, 0.01200, 0.01400	0.00800, 0.01100, 0.01400	0.00600, 0.00900, 0.01300	0.00300, 0.00600, 0.00900	0.00200, 0.00500, 0.00700	0.00200, 0.00400, 0.00500	0.00400, 0.00600, 0.00700	0.00300, 0.00400, 0.00600	0.00800, 0.01100, 0.01300
DC5	0.00400, 0.00700, 0.01000	0.00800, 0.01100, 0.01400	0.00600, 0.00900, 0.01300	0.01100, 0.01400, 0.01600	0.00800, 0.01100, 0.01300	0.00200, 0.00400, 0.00500	0.00400, 0.00500, 0.00700	0.00000, 0.00100, 0.00300	0.00500, 0.00800, 0.01100
DC6	0.02500, 0.03900, 0.05300	0.04100, 0.05300, 0.05800	0.05200, 0.06600, 0.07400	0.03700, 0.04700, 0.05200	0.02100, 0.02900, 0.03700	0.01900, 0.02600, 0.03200	0.03500, 0.04900, 0.06300	0.02700, 0.04100, 0.05600	0.01600, 0.02400, 0.03200
DC7	0.02100, 0.03500, 0.04900	0.02900, 0.04100, 0.05300	0.04100, 0.05600, 0.06900	0.03300, 0.04300, 0.05000	0.01200, 0.02100, 0.02900	0.01000, 0.01700, 0.02400	0.02900, 0.04001, 0.05300	0.01900, 0.02900, 0.04000	0.01200, 0.01900, 0.02600

TABLE. VII. CLOSENESS COEFFICIENTS TO ASPIRED LEVEL AMONG DIFFERENTALTERNATIVES

	dþi	Di	Satisfaction degree of CCi	Ranks
HB1	0.7400	29.1000	0.42300	5
HB2	0.7100	29.2000	0.52410	2
HB3	0.7200	29.3000	0.42200	7
HB4	0.7300	29.4000	0.52400	3
DC5	0.6500	29.0000	0.62220	6
DC6	0.6900	29.0012	0.62312	4
DC7	0.6600	29.1240	0.52431	1

VI. Comparison through Classical ANP-TOPSIS And FUZZY ANP-TOPSIS Method

In this paper, authors used classical AHP-TOPSIS technique and FUZZY AHP-TOPSIS for comparison[21] to verifying the accuracy of the results. In Fuzzy and classical AHP-TOPSIS, both have the same techniques to collect and assessment data[21]. No fuzzification required in classical AHP this is the main difference with Fuzzy AHP. In classical AHP-TOPSIS, data is taken in numeric form. The results difference between fuzzy and classical AHP-TOPSIS is shown in table 8. The obtained result by the classical AHP-TOPSIS method and fuzzy AHP-TOPSIS method is highly correlated. The accuracy of Fuzzy AHP TOPSIS is better than the classical AHP TOPSIS shown in results.

TABLE. VIII. Comparison the Results of Classical and FUZZY AHP-TOPSIS Methods

Alternatives	Fuzzy ANP-TOPSIS	Classical ANP-TOPSIS
HB1	0.42300	0.41100
HB2	0.52410	0.51460
HB3	0.42200	0.40950
HB4	0.52400	0.51300
DC5	0.62220	0.61320
DC6	0.62312	0.62162
DC7	0.52431	0.51832

VII. Discussion

Shocking increase in breaches of medical image has been seen by hospitals recently. In the first half of 2019 itself, 32 million health records were breached, in comparison to 15 million in the whole year of 2018. Patients' health information like date of birth, medical history, credit/debit card number and other classified details can be manipulated, corrupted and worst, sold for a high price in the market. Dismally, only 4% to 7% of revenue is invested by the healthcare industry in security. According to the Verizon Data Breach Investigation report, main source of security breaches in healthcare are insiders. The report states that 59% of the breaches in 2018 were done by insiders and 42% by external invaders. For positing an efficacious solution to this anomaly, the researchers of the present study have proposed the Fuzzy AHP-TOPSIS for security assessment. The study places an empirical evidence to suggest that affectivity of the security condition given to medical image processing system can be gauged by this methodology. The systems that are chosen for this study are being used by the hospitals and diagnostic laboratories in Varanasi. To protect the privacy of these healthcare centers, we have not enlisted their identity in this research. After collating the data from these avenues and assimilating the feedback of the practitioners about the contribution of security of medical image processing system at the time of processing, the information collected from the experts is calculated through the Fuzzy AHP- TOPSIS approach. Findings can be précised as:

- Assessment of the security of medical image will help the developers to focus on users' satisfaction.
- Through the Fuzzy AHP- TOPSIS we get the quantitative results that will support in categorizing the higher ranked factor for security assessment while developing the system.
- Development guidelines produced through this estimation will help the developers to improve their products and aid the government organizations in checking the project in the pre-market.

It is clear from this discussion that security assessment needs inventive methodologies that must be workable and accurate. Our research has worked on Fuzzy AHP- TOPSIS, yet the future challenges to reckon are:

- We have used this approach in web based medical image processing security. May be some security attributes have been missed in our empirical analysis.
- Results may be changed if the weights of inputs changed in FAHP.

VIII. Conclusion

The software industry has developed insecure system with various vulnerabilities which are non-acceptable in the medical field. This paper tests seven projects and presents a comprehensive study of security assessment of medical image processing system. The results of our tests show that most of the systems are at risks and they need to improve the security. Manually assessment of the security in medical image processing system is difficult. Proposed framework provides the quantitative assessment of security in the terms of ranking of the system. Healthcare industry does not want to invest revenue in the security. This framework will reduce the cost and time spent in the security checking. A list of criteria and Fuzzy AHP-TOPSIS methods provide rank to the system according to security check. This system provides a development guideline to improve the security at the time of software development. Thus this assessment of security will help the government and software industry to develop guidelines to make medical image processing system/tool more secure. Future challenge in our work, this technique is based on weight selection if any weight can changed by default then results can also changed.

References

[1] Jagannathan, Srinivasan&Sorini, Adam. (2015). A cybersecurity risk analysis methodology for medical devices. 1-6.

[2] Marwan, Mbarek&Kartit, Ali &Ouahmane, Hassan. (2018). A Cloud-based Framework to Secure Medical Image Processing. Journal of Mobile Multimedia. 14. 319-344.

[3] Arun, Ashish& George (2018). Healthcare Informatics and Privacy.

[4] Singh, A. & Chatterjee, K. 4ITrust: identity and trust based access control model for healthcare system security(2019) 78: 28309. https://doi.org/10.1007/s11042-019-07923.

[5] Rong, Ming &Zhi, Shuyue&Peng, Qingguo. (2016). Rayplus: A web-based platform for Medical Image Processing.

[6] Natsheh, Qamar& Li, Baihua& Gale, Alastair. (2016). Security of Multi-frame DICOM Images Using XOR Encryption Approach. Procedia Computer Science. 90. 175-181.

[7] Yinghui, Pengzhen, Dong, Menglei, &Rui (2018). A secure and privacy-Aware Smart Health system with Secret Key Leakage Resilience.

[8] Viadimir, &Marios (2019). Security and Privacy Of Medical Data: Challenges for next generation Patient-Centric Healthcare Systems.

[9] Riyaz Belgaum, Mohammad &Alansari, Zainab& Jain, Ruchin&Alshaer, Jawdat. (2018). A Framework for Evaluation of Cyber Security Challenges in Smart Cities.

[10] Bansal, Siddhant& Mehta, Garima. (2017). Comparative analysis of joint encryption and watermarking algorithms for security of biomedical images. 609-612.

[11] Wang, Zhiqiang& Ma, Pingchuan& Chi, Yaping& Zhang, Jianyi. (2018). Medical Devices are at Risk: Information Security on Diagnostic Imaging System. 2309-2311.

[12] Shanmugapriya&Kavitha (2019). Medical big data analysis: preserving security and privacy with cloud technology.

[13] Modi K.J., Kapadia N. (2019) Securing Healthcare Information over Cloud Using Hybrid Approach. In: Panigrahi C., Pujari A., Misra S., Pati B., Li KC. (eds) Progress in Advanced Computing and Intelligent Engineering. Advances in Intelligent Systems and Computing, vol 714. Springer, Singapore.

[14] Ak, MuhammetFatih&Gül, Muhammet. (2018). AHP–TOPSIS integration extended with Pythagorean fuzzy sets for information security risk analysis. Complex & Intelligent Systems.

[15] Shi, M., Jiang, R., Hu, X. et al. A privacy protection method for health care big data management based on risk access control(2019).

[16] Aqsa, & Ricardo (2018). Security aspects in healthcare information systems: A systematic mapping.

[17] Ma, Pingchuan& Wang, Zhiqiang&Hei, Xiali&Zou, Xiaoxiang& Zhang, Jianyi& Liu, Qixu&Lyu, Xin&Zhuo, Zihan. (2019). A Quantitative Approach for Medical Imaging Device Security Assessment. 5-6.

[18] A. Agrawal, M. Alenezi, R. Kumar and R. A. Khan, "Measuring the Sustainable-Security of Web Applications Through a Fuzzy-Based Integrated Approach of AHP and TOPSIS," in *IEEE Access*, vol. 7, pp. 153936-153951, 2019.

[19] Karim, Abderrahim& Hayat (2018). Big healthcare data: preserving security privacy.

[20] Chang DY. Applications of the extent analysis method on fuzzy AHP. Eur J Oper Research. 1996; 95:649–55.

[21] Moayeri, M. & Shahvarani, A. & Behzadi, M.H. & Hosseinzadeh-Lotfi, F.. (2015). Comparison of fuzzy AHP and fuzzy TOPSIS methods for math teachers selection. Indian Journal of Science and Technology. 8. 10.17485/ijst/2015/v8i13/54100.

Knowledge Sharing Factors for Modern Code Review to Minimize Software Engineering Waste

Nargis Fatima[1], Sumaira Nazir[2], Suriayati Chuprat[3]
Razak Faculty of Technology and Informatics, University Technology Malaysia (UTM), Kuala Lumpur, Malaysia [1, 2, 3]
Department of Engineering and Computer Science, National University of Modern Languages, (NUML), Islamabad, Pakistan[1, 2]

***Abstract*—Software engineering activities, for instance, Modern Code Review (MCR) produce quality software by identifying the defects from the code. It involves social coding and provides ample opportunities to share knowledge among MCR team members. However, the MCR team is confronted with the issue of waiting waste due to poor knowledge sharing among MCR team members. As a result, it delays the project delays and increases mental distress. To minimize the waiting waste, this study aims to identify knowledge sharing factors that impact knowledge sharing in MCR. The methodology employed for this study is a systematic literature review to identify knowledge sharing factors, data coding with continual comparison and memoing techniques of grounded theory to produce a unique and categorized list of factors influencing knowledge sharing. The identified factors were then assessed through expert panel for its naming, expressions, and categorization. The study finding reported 22 factors grouped into 5 broad categories i.e. Individual, Team, Social, Facility conditions, and Artifact. The study is useful for researchers to extend the research and for the MCR team to consider these factors to enhance knowledge sharing and to minimize waiting waste.**

Keywords—Knowledge sharing; modern code review; software engineering waiting waste

I. Introduction

Software engineering is a socio-technical activity for the development of software with specified resources [1]. It includes activities such as requirement identification, modeling, construction, testing, and Modern Code Review (MCR) [2]. These activities produce various wastes such as waiting, development of extra or erroneous feature, defect, needless composite solution, rework, and mental distress [2], [3], [4], [5]. In software engineering, waste can be defined as "anything that doesn't make it to the release, is waste" [4].

Modern Code Review, a lightweight form of traditional Fagan's code inspection [6], has been expanding in the research. A Fagan examination is a heavyweight code inspection procedure requiring synchronous interactions among the members in multiple stages [7]. On the other hand, MCR is characterized as being trivial, increasingly casual, and strengthened by review tools [6], [8], [9]. Notwithstanding studies that confirm Fagan's code inspections advances the quality of software [7], [10] their required cost and formality have prohibited widespread acceptance [6], [8], [9]. Contrariwise, MCR has addressed many inadequacies of Fagan's code inspection and highly adopted in industry and open-source software development contexts [6], [8], [9], [11].

Although MCR has addressed many shortcomings of Fagan's code inspections and is developed to improve software and code quality through extensive knowledge sharing among MCR team members [6], [8], [9], [11], [12], however, the MCR generates waiting waste due to poor knowledge sharing [5], [8], [13], [14], [15], [16], [17], [18].

Current researchers [8], [9], [17] have shown that MCR team members are hesitant to share knowledge and give a timely response to other members and let them in a waiting condition. It is argued that waiting waste can be minimized by increasing knowledge sharing [2], [4], [5], [19] among the MCR team. It is also argued that knowledge sharing can be increased by identifying the factors influencing knowledge sharing [8], [9], [11], [20], [21] that can increase knowledge sharing among the MCR team that might lead to the reduction in the production of waiting waste in MCR.

Although previous research has given attention to knowledge sharing concerning software engineering activities [22], [23], [24], [25], however, knowledge sharing in the context of MCR has not got much attention from the researchers [8], [9], [11], [20], [21]. No, systematize investigations are available concerning the knowledge sharing aspect in MCR that can help in minimizing waiting waste. Therefore, the purpose of this study is to perform a Systematic Literature Review (SLR) to produce a validated and unique list of factors influencing knowledge sharing in MCR to minimize waiting waste.

The rest of this paper is distributed as Section II describes the background and related work. Section III covers the search method while Section IV introduces the results of SLR and expert review. Section V provides the discussion; Section VI presents the limitation of the study. Section VII presents the conclusion. Section VIII provides future directions. Section IX highlights the contribution of the study.

II. Background

Software engineering is a development of quality software within a stated time and budget [1]. The success factor of software engineering is subject to whether the software can fulfill user demands [1]. Software engineering is a socio-technical activity that incorporates managing other activities [2], [5] such as requirement identification, modeling, construction, testing, and Modern Code Review (MCR). These activities deliver ample prospects of producing wastes [2], [4], [5]. Waste is any act that does not produce any value to the user [2]. Concerning software engineering it can be "anything that doesn't make it to the release, is waste" [4]. It can also refer to any activity which uses resources but does not produce quality software [2], [4].

MCR is a software engineering activity for code improvement [6]. In MCR the code is reviewed by the reviewer, before committing the code to the project codebase. Unlike Fagan formal inspection process, MCR focuses on reviewing the small part of code changes usually named as 'patch' before saving the code into the codebase [26]. MCR regularly occurs in practice [8], [9], [11] with the help of code review tools [6], [9] such as Gerrit, Code flow, Review board, Phabricator, etc. It is a means to identify defects and to improve code quality [2], [6], [8], [9], [11], [12], [27], [28], [29] through knowledge sharing among developers. Fig. 1 shows the MCR process overview.

It is argued that MCR produces wastes such as waiting, development of extra or erroneous feature, defect, needless composite solution, rework, and mental distress [2], [3], [4], [5]. It is contended that waiting is the critical wastes [4], [30], [31]. It is argued that "one of the biggest wastes in software development is usually waiting for the thing to happen"[30]. It is also conveyed that if the organization has to minimize one waste, it should focus on a waiting [4], [30], [31].

Waiting waste refers to a delay between two consecutive activities [2], [3], [4], [5], [30], [31]. For example, in MCR delay between submitting source code review request by the author to the reviewer and getting feedback from the reviewer [8], [9], [17]. It is argued that one of the reasons for waiting waste in MCR is a poor knowledge sharing [5], [8], [13], [14], [15], [16], [17], [18]. The waiting waste decreases the productivity and efficiency of the developers [2], [4], [8], [12], [16], [17], [21], [26], [32]. It also causes project delays due to the blocking of tasks [4], [33].

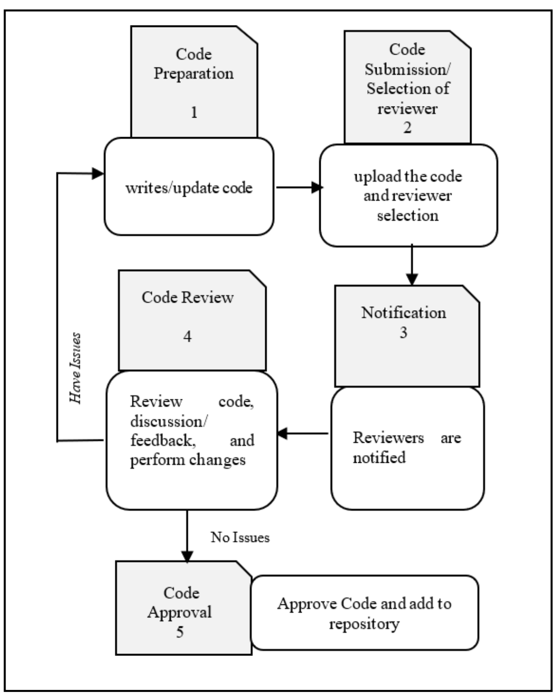


Fig. 1. MCR Process Overview [9].

To minimize the waiting waste it is necessary to increase knowledge sharing [2], [4], [5], [19] among MCR team members. It is argued that knowledge sharing can be increased by identifying the factors influencing knowledge sharing [8], [9], [11], [20], [21] that can help in effective knowledge sharing among the MCR team.

Though preceding studies [22], [23], [24], [25] focused on knowledge sharing in software engineering activities, however, slight indication is available in MCR [8], [9], [11], [20], [21], resulting in absence of knowledge sharing guidelines in MCR. Therefore, the study aims to identify factors influencing knowledge sharing in MCR to minimize waiting waste.

Systematic Literature Review (SLR), has been directed to identify the factors influencing knowledge sharing in MCR. The expert review has been performed to confirm the identified factors influencing knowledge sharing for their naming, expressions, and categories.

III. Research Methodology

Multiple research activities have been performed to generate a distinct and categorized rundown of factors influencing the knowledge sharing in MCR to minimize the generation of waiting waste. The methodologies employed for this study are discussed in subsections.

A. *Systemantic Literature Review*

The Systematic Literature Review (SLR) methodology given by [34] has been used for this study to identify the relevant data sources for the identification of factors influencing knowledge sharing in MCR to minimize the generation of waiting waste. The SLR methodology is a systematized and well-organized approach to attain less impartial results [34]. It is an authentic methodology to record significant central focuses in the research for assessing and looking at all momentum research related to research questions. The detailed procedure of SLR is explained in subsections.

1) Research question: Constructing the research question is the central action of SLR [34]. Research questions are designed with the support of PICOC criteria specified by Petticrew and Roberts [35]. This investigation has excluded the 'Comparison' segment of the PICOC yet just PIOC has been considered to design the research question. The reason behind excluding the comparison part is that this study is not considering the comparison of techniques or models. Table I represents the PIOC criteria for this study.

TABLE. I. POIC Summary

Population	MCR team
Intervention	MCR Process
Outcome	Factors influencing knowledge sharing in MCR to minimize waiting waste.
Context	The study includes all study types such as interviews, observations, surveys, experiments, questionnaires and case studies relating to MCR.

To gather the indications on the present state of research regarding factors influencing knowledge sharing in MCR to reduce waiting wastes. The designed question is specified below.

RQ1: What factors influence the knowledge sharing in MCR to minimize software engineering waiting waste?

2) Search Strategy: The search strategy comprises of identification of key terms and their alternate substitutes.

a) Identification of key term: The study key terms includes knowledge sharing, modern code review and software engineering waiting waste

b) Finding substitutes of identified key terms: The substitutes for the identified key terms are shown in Table II.

c) Use of Boolean OR to design search strings with key terms and their substitutes: The key terms along with their substitutes are joined using Boolean OR and are represented in Table III.

d) Use Boolean AND to concatenate the search key terms and limit the research: The designed search string is given below.

(‘Knowledge sharing’ OR ‘knowledge distribution’ OR ‘knowledge transfer’, ‘knowledge dissemination’ OR ‘knowledge exchange’) AND (review’ OR ‘modern code inspection ’OR ‘code review’ OR ‘code inspection ’OR ‘lightweight code review’) AND (‘Software Engineering Waiting Waste’ OR ‘software engineering delay waste’ OR ‘software engineering linger waste’ OR ‘software engineering blocking waste’ OR ‘software development delay waste’ OR ‘software development linger waste’)

TABLE. II. KEY TERMS AND THEIR SUBSTITUTES

Key term	Substitutes
Knowledge sharing	‘knowledge distribution’, ‘knowledge transfer’, ‘knowledge dissemination’, ‘knowledge exchange’
Modern Code Review	‘contemporary code review’, ‘modern code inspection’, ‘code review’, ‘code inspection’, ‘lightweight code review’
Software Engineering Waiting Waste	‘software engineering delay waste’, ‘software engineering linger waste’, ‘software engineering blocking waste’, ‘software development delay waste’, ‘software development linger waste’

TABLE. III. KEY TERMS WITH THEIR SUBSTITUTES AND BOOLEAN OR OPERATOR

Key terms, Substitutes and Boolean OR
‘Knowledge sharing’ OR ‘knowledge distribution’ OR ‘knowledge transfer’, ‘knowledge dissemination’ OR ‘knowledge exchange’
‘Modern Code Review’ OR ‘contemporary code review’ OR ‘modern code inspection ’OR ‘code review’ OR ‘code inspection ’OR ‘lightweight code review’
‘Software Engineering Waiting Waste’ OR ‘software engineering delay waste’ OR ‘software engineering linger waste’ OR ‘software engineering blocking waste’ OR ‘software development delay waste’ OR ‘software development linger waste’

e) Search process and database sources: The search process involved databases such as IEEE, Science Direct, ACM, Wiley online, Springer link, Web of Science and Scopus. The reason for selecting the above databases is that the selected databases are known to have software engineering literature. To make the search process comprehensive and to avoid the chance of missing out evidence, the search included the literature published from 2013 – 2019. Database sources that were considered are presented in Table IV along with their URLs and distribution.

f) Study Selection Criteria: The studies are included and excluded based on the inclusion and exclusion plan shown in Fig. 2.

Study Quality Assessment: Notwithstanding broad inclusion and exclusion criteria, it is viewed as basic to evaluate the "quality" of essential investigations. For the evaluation of concentrate quality, the checklist specified by [34] has been used. The investigations chosen after the introductory inclusion and exclusion plan are additionally assessed utilizing the checklist articulated in Table V.

The questions specified in the checklists represented in Table V are answered according to the rule specified by [34]. The evaluation scale is presented in Table VI.

TABLE. IV. DATABASE SOURCES

Data Source	URL
IEEE	http://ieeexplore.ieee.org/
ACM	http://dl.acm.org
Science Direct	http://www.sciencedirect.com
Wiley	http://onlinelibrary.wiley.com
Web of Science	https://www.webofknowledge.com
Springer link	https://www.springer.com
Scopus	http://www.scopus.com

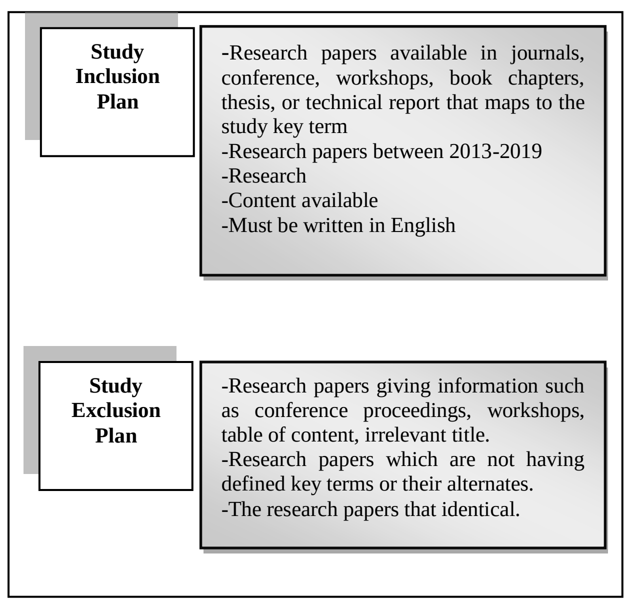


Fig. 2. INCLUSION AND EXCLUSION PLAN.

TABLE. V. QUALITY ASSESSMENT CHECKLIST

Question	**Answer**
Are the goals visibly detailed?	Yes/ No/Partially
Are the outcomes complete and substantial?	-do-
Are the prediction methods used visibly defined and their choice are acceptable?	-do-
Is the information been extended by the study?	-do-
Is the diversity of viewpoint and context been sightseen?	-do-
Are the links between data, understanding, and assumptions are vibrant?	-do-
Does the difficulty of the data is transferred?	-do-

TABLE. VI. SCALE FOR ANSWERING QUESTIONS GIVEN IN CHECKLIST [34]

Answer	**Score**
Yes	1
No	0
Partially	0.5

g) Data Extraction: After the essential studies have been chosen and their quality assessed, the data is extracted from the selected papers. The data extraction method is discussed in this section. The data extraction method is intended to contain all the data that is important for responding to the research question and tending to the investigation quality criteria [34]. The data extraction form is represented in Table VII.

h) Data Synthesis: After vigilant data extraction the extracted data is synthesized following the data coding, continual data comparisons and memoing from grounded theory [36] are adopted for data unit categorization, and to get the unique list of factors influencing knowledge sharing in MCR.

B. *Expert Review*

After getting the unique list of factors influencing knowledge sharing in MCR the list is evaluated through experts for naming, expression, categorization, and suggestions of new factors or categories. The considered experience for experts' selection is more than 10 years in software development knowing MCR, software engineering wastes and knowledge sharing. For expert review, the guidelines of Ayyub [30] are followed.

TABLE. VII. DATA EXTRACTION FORM [34]

Data characteristics	A unique identifier in the format: KSFP(1)…KSFP(n)
	Title
	Author (s)
	Year
	Study Set (Conference/Journal)
	Study Commissioner (IEEE, ACM, etc.)
	Selection (Inclusion/exclusion)/Quality assessment
Research Question	What factors influencing knowledge sharing in MCR?

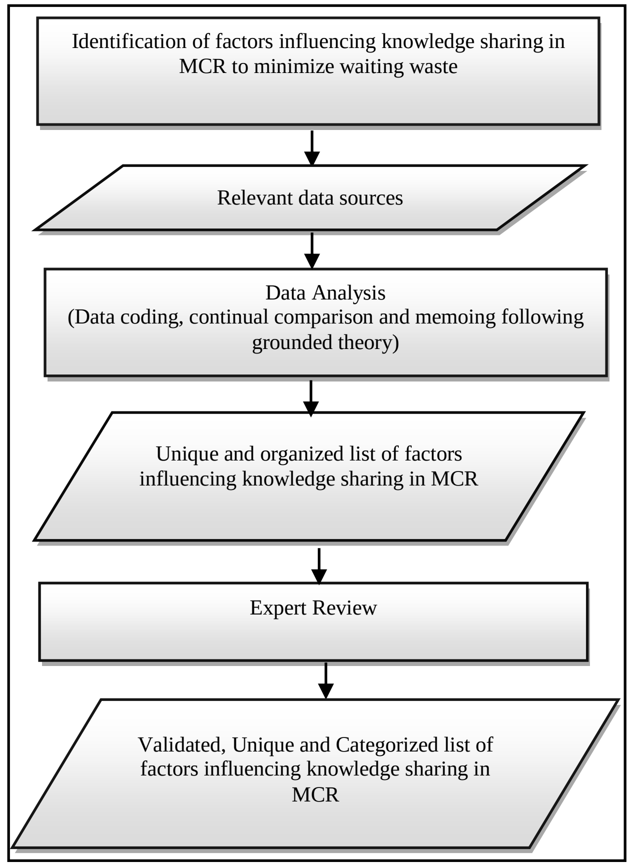


Fig. 3. DATA SYNTHESIS PROCEDURE.

Fig. 3 summarizes the data synthesis procedure employed for this study.

IV. RESULTS

This section discusses the results achieved in the study. It presents the results concerning the study search process to achieve pertinent data sources and the factors influencing knowledge sharing in MCR to minimize waiting waste.

A. *Data Source Selection Results*

Through initial search based on defined key terms, 9289 papers are obtained. The studies that represent only the table of content, conference or workshop preceding details or having unrelated titles are omitted. After the first exclusion, 1103 studies are obtained. The obtained 1103 studies are evaluated for the relevant key terms (modern code review, knowledge sharing, and software engineering waiting waste). The studies that do not have any of the correlated key term are eliminated and 190 studies are included. After assessment for having duplication among 190 studies, 162 studies are obtained and evaluated for their quality assessment. During the quality assessment, 6 studies are excluded and finally, 156 studies are recognized as most appropriate to this study and are included for detailed review.

B. *Knowledge Sharing Factors in MCR*

This section stretches the insights about factors influencing knowledge sharing in MCR to minimize waiting waste. The study results reported 22 factors that impact knowledge sharing in MCR, the identified factors are grouped under 5 broad categories namely Individual, Team, Social, Artifact and Facility Conditions. The details are represented in subsections. Table VIII provides a summarized view of the factors influencing knowledge sharing in MCR along with their references.

1) Individual: Individual perspective is the most noticeable lens in MCR [32]. The factors involved in this category are individual impartiality, individual historical factors, individual emotions, individual pressure, individual awareness, individual turnover, and individual intentions [9], [11], [17], [19], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47].

2) Social: MCR is a multifaceted process that involves social interactions among team members [32]. This category includes factors i.e. relational and structural factors [8], [9], [11], [17], [19], [48], [49], [50], [51].

3) Team: The team signifies a group of individuals who worked together to achieve a common goal. Their work involved multiple projects, from new to legacy systems [8]. This category involves factors i.e. team organization, team strategies, team culture, team, and team drive [8], [9], [11], [16], [32], [52], [48], [40], [53].

4) Artifact: An artifact, is an object made or given form by humans [12], [32]. This category includes factors such as source code, testing, feedback [8], [9], [11], [16], [19],[32], [52], [54], [55], [56], [57], [58], [59], [60].

5) Facility Conditions: Facility conditions support the successful conduction of the MCR [32]. This category involves factors i.e. project, process, tool, communication channel, and organization [9], [11], [19], [12], [15], [32].

Table VIII summarizes the validated list of factors influencing knowledge sharing in MCR prompting knowledge sharing in MCR along with the references.

TABLE. VIII. LIST OF KNOWLEDGE SHARING FACTORS IN MCR TO MINIMIZE SOFTWARE ENGINEERING WASTES

Categories	**Knowledge Sharing Factors**	**References**
INDIVIDUAL	Individual Impartiality	[9], [11], [17]
	Individual historical factors	[6], [8], [9], [16], [11], [17], [19], [32], [37], [52], [45], [48], [61], [62], [49], [38], [39], [40] , [41], [42] , [43], [44], [46]
	Individual Emotions	[8], [9], [15], [17], [32], [52] [63], [64]
	Individual Pressure	[6], [8], [9], [11], [15],[19], [32], [52], [48], [49], [40], [65], [54],
	Individual Awareness	[8], [9], [11], [14], [19], [32], [37], [52], [48], [49], [44], [65], [54], [66], [55], [56], [67],
	Individual Turnover	[64]
	Individual Intentions	[9], [11], [12], [17], [19], [37], [52], [61], [49] [64], [54], [56], [68], [69], [70],
SOCIAL	Relational	[8], [9], [11], [16], [17], [19], [32] , [48], [61], [49], [39], [40], [41], [42], [43], [44], [54], [70], [37], [57], [71], [72], [73], [74], [75]
	Structural	[15], [44], [50], [51]
ARTIFACT	Source Code	[8], [9], [11], [16], [19],[32], [6], [12] [15], [52], [45], [48], [38], [40], [41], [46], [63], [65], [54], [55], [56], [76], [77], [58], [78], [79]
	Feedback	[8], [9], [11], [15], [19], [32] [48], [40], [63], [54], [55], [56], [57], [72], [76], [58], [59]
	Testing	[8], [9], [11], [15], [19], [32], [52],[48], [75], [58], [59], [60],
FACILITY CONDITIONS	Process	[8], [9], [11], [19], [52], [48], [39], [78]
	Tool	[6], [8], [11], [12], [15], [32], [38] , [55], [71] [77], [78]
	Organization	[8], [12], [17], [32], [52], [38]
	Communication	[8], [9], [15], [52], [48], [38], [55]
	Project	[9], [11], [15], [32], [48]
TEAM	Team Organization	[8], [9], [11], [16], [32]
	Team Strategies	[8], [12], [15], [52]
	Team Culture	[8], [11], [52]
	Team Intensions	[6], [8], [9], [12], [32], [48], [40], [56]
	Team Drive	[8], [9], [11], [19], [32], [52], [48], [40], [53]

V. Discussion

This work stretches the direction to a comprehensive list of factors influencing knowledge sharing in MCR to minimize waiting waste. The identified factors are significant for software engineers involved in the MCR process. The preliminary list can act as a guide for the researchers and practitioners working in MCR to consider and these factors in order to increase knowledge sharing and to minimize waiting waste. This study contributed to the software engineering body of knowledge (SWEBOK) particularly to knowledge sharing in the context of MCR. The study helps the MCR team to achieve its objective while minimizing waiting waste.

VI. Limitations

This study lacks the identification of factors from the industry as the study comprises of factors that are recognized from the literature. A large effort has been made to cover all the correlated papers, but still, there is a possibility that some research may be missed.

VII. Conclusion

The research study provides a categorized list of factors influencing knowledge sharing in MCR to minimize waiting waste. The reported factors that influence knowledge sharing in MCR are distributed into five main categories that are Individual, Social, Team, Artifact and Facility Conditions. These factors ought to be considered while performing MCR to minimize waiting waste by increasing knowledge sharing.

VIII. Future Directions

A comprehensive list will be produced in the future by quantitative analysis, the ongoing research objectives. In addition to this, a comprehensive model can be produced for MCR that can be used as a guideline for software engineers to minimize software engineering waiting waste. This work recognizes factors influencing knowledge sharing in MCR that provides the foundation for the investigators to outspread this research by discovering other factors for other software development activities to reduce wastes.

IX. Contribution

The examination contributed towards software engineering body of knowledge (SWEBOK), knowledge base software engineering (KBSE) and green software engineering (GREEN SE) by perceiving the significance of knowledge sharing and by giving the arranged rundown of factors influencing knowledge sharing in MCR. The work can help software developers to successfully transfer knowledge by overcoming the negative aspects of identified factors.

References

[1] Alvertis et al., “User Involvement in Software Development Processes,” Procedia Comput. Sci., vol. 97, pp. 73–83, 2016.

[2] T. Sedano and P. Ralph, “Software Development Waste,” in Proc. IEEE/ACM 39th International Conference on Software Engineering, 2017.

[3] M. Poppendieck and T. Poppendieck, “Implementing Lean Software Development: From Concept to Cash,” Addison-Wesley Signat. Ser., p. 304, 2006.

[4] H. Alahyari, T. Gorschek, and R. Berntsson Svensson, “An exploratory study of waste in software development organizations using agile or lean approaches: A multiple case study at 14 organizations,” Inf. Softw. Technol., vol. 105, no. August 2018, pp. 78–94, 2019.

[5] “Software Engineering Wastes – A Perspective of Modern Code Review,” 2020.

[6] A. Bacchelli and C. Bird, “Expectations, outcomes, and challenges of modern code review,” in Proc. International Conference on Software Engineering, 2013, pp. 712–721.

[7] M. E. Fagan, “Design and code inspections to reduce errors in program development,” IBM Syst. J., vol. 38, no. 2.3, pp. 258–287, 1999.

[8] L. MacLeod, M. Greiler, M. A. Storey, C. Bird, and J. Czerwonka, “Code Reviewing in the Trenches: Challenges and Best Practices,” IEEE Softw., vol. 35, no. 4, pp. 34–42, 2018.

[9] A. Bosu, J. C. Carver, C. Bird, J. Orbeck, and C. Chockley, “Process Aspects and Social Dynamics of Contemporary Code Review: Insights from Open Source Development and Industrial Practice at Microsoft,” IEEE Trans. Softw. Eng., vol. 43, no. 1, pp. 56–75, 2017.

[10] A. Aurum, H. Petersson, and C. Wohlin, “State-of-the-art: Software inspections after 25 years,” Softw. Test. Verif. Reliab., vol. 12, no. 3, pp. 133–154, 2002.

[11] C. Sadowski, E. Söderberg, L. Church, M. Sipko, and A. Bacchelli, “Modern code review: : A Case Study at Google,” in Proc. ACM/IEEE 40th International Conference on Software Engineering: Software Engineering in Practice, 2018, pp. 181–190.

[12] P. C. Rigby and C. Bird, “Convergent Contemporary Software Peer Review Practices Categories and Subject Descriptors,” in Proc. ESEC/FSE, 2013, pp. 202–212.

[13] J. C. Carver, B. Caglayan, M. Habayeb, B. Penzenstadler, and A. Yamashita, “Collaborations and code reviews,” IEEE Softw., vol. 32, no. 5, pp. 27–29, 2015.

[14] J. Czerwonka, M. Greiler, and J. Tilford, “Code Reviews Do Not Find Bugs. How the Current Code Review Best Practice Slows Us Down,” in Proc. IEEE/ACM 37th IEEE International Conference on Software Engineering, 2015, vol. 2, pp. 27–28.

[15] G. Gousios, M.-A. Storey, and A. Bacchelli, “Work practices and challenges in pull-based development,” in Proc. 38th International Conference on Software Engineering, 2016, pp. 285–296.

[16] E. W. dos Santos and I. Nunes, “Investigating the Effectiveness of Peer Code Review in Distributed Software Development,” in Proc. 31st Brazilian Symposium on Software Engineering, 2017, pp. 84–93.

[17] D. M. German, U. Rey, and J. Carlos, “‘ Was my contribution fairly reviewed ?’ A Framework to Study the Perception of Fairness in Modern Code Reviews,” in Proc. ACM/IEEE 40th International Conference on Software Engineering Synthesizing, 2018, no. 2, pp. 523–534.

[18] L. Novikova, “Poor knowledge sharing is the second biggest challenge for software development teams,” 2019. [Online]. Available: https://blog.onebar.io/poor-knowledge-sharing-is-the-second-biggest-challenge-for-software-development-teams-a4843f9b9aa. [Accessed: 10-Aug-2019].

[19] A. Ram, Achyudh ; Sawant, Anand; Castelluccio, Marco; Bacchelli, “What Makes a Code Change Easier to Review? An Empirical Investigation on Code Change Reviewability,” in Proc. ESEC/FSE, 2018.

[20] N. Fatima, S. Nazir, and S. Chuprat, “Knowledge sharing, a key sustainable practice is on risk: An insight from Modern Code Review,” in 2019 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS), 2019.

[21] N. Fatima, S. Nazir, and S. Chuprat, “Understanding the Impact of Feedback on Knowledge Sharing in Modern Code Review,” in 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) , In Press, 2019.

[22] R. Anwar et al., “Conceptual Framework for Implementation of Knowledge Sharing in Global Software Development Organizations,” pp. 174–178, 2017.

[23] X. Chen, Y. Zhou, D. Probert, and J. Su, “Technology Analysis & Strategic Management Managing knowledge sharing in distributed innovation from the perspective of developers : empirical study of open source software projects in China,” Technol. Anal. Strateg. Manag., vol. 7325, no. July, 2016.

[24] N. S. Safa and R. Von Solms, “An information security knowledge sharing model in organizations,” Comput. Human Behav., vol. 57, pp. 442–451, 2016.

[25] Q. Zhang and R. Du, “Impacts of cultural difference on knowledge sharing, relationship quality and performance in IT-based service outsourcing,” 2011 2nd Int. Conf. Artif. Intell. Manag. Sci. Electron. Commer. AIMSEC 2011 - Proc., pp. 6271–6274, 2011.

[26] C. Thompson and D. Wagner, “A Large-Scale Study of Modern Code Review and Security in Open Source Projects,” Proc. 13th Int. Conf. Predict. Model. Data Anal. Softw. Eng., pp. 83–92, 2017.

[27] S. McIntosh, Y. Kamei, B. Adams, and A. E. Hassan, “An empirical study of the impact of modern code review practices on software quality,” Empir. Softw. Eng., vol. 21, no. 5, pp. 2146–2189, 2016.

[28] N. Fatima, S. Chuprat, and S. Nazir, “Challenges and Benefits of Modern Code Review-Systematic Literature Review Protocol,” in Proc. International Conference on Smart Computing and Electronic Enterprise, 2018, pp. 1–5.

[29] S. Nazir, N. Fatima, and S. Chuprat, “Modern Code Review Benefits– Primary findings of a systematic literature review,” in The 3rd International Conference on Software Engineering and Information Management.

[30] M. Poppendieck and T. Poppendieck, Lean Software Development: An Agile Toolkit. 2003.

[31] J. Urrego, R. Munoz, M. Mercado, and D. Correal, “Archinotes: A global agile architecture design approach,” Lect. Notes Bus. Inf. Process., vol. 179 LNBIP, pp. 302–311, 2014.

[32] O. Kononenko, O. Baysal, and M. W. Godfrey, “Code Review Quality: How Developers See It,” in Proc. International Conference on Software Engineering, 2016, pp. 1028–1038.

[33] M. Ikonen, P. Kettunen, N. Oza, and P. Abrahamsson, “Exploring the sources of waste in Kanban software development projects,” in Proc. - 36th EUROMICRO Conference on Software Engineering and Advanced Applications, 2010, pp. 376–381.

[34] B. Kitchenham and S. Charters, “Source: " Guidelines for performing Systematic Literature Reviews in SE " , Kitchenham et al Guidelines for performing Systematic Literature Reviews in Software Engineering Source: " Guidelines for performing Systematic Literature Reviews i,” pp. 1–44, 2007.

[35] F. Terms, “ M. Petticrew and H. Roberts. Systematic Reviews in the Social Sciences: A Practical Guide . Oxford: Blackwell 2006. 352 pp. ISBN 1 4051 2110 6. £29.99 ,” Couns. Psychother. Res., vol. 6, no. 4, pp. 304–305, 2006.

[36] K.-J. Stol, P. Ralph, and B. Fitzgerald, “Grounded theory in software engineering research,” no. October 2017, pp. 120–131, 2016.

[37] A. Bosu, J. Carver, R. Guadagno, B. Bassett, D. McCallum, and L. Hochstein, “Peer impressions in open source organizations: A survey,” J. Syst. Softw., vol. 94, pp. 4–15, 2014.

[38] F. Ebert, F. Castor, N. Novielli, and A. Serebrenik, “Confusion detection in code reviews,” in Proc. IEEE International Conference on Software Maintenance and Evolution, 2017, pp. 549–553.

[39] A. Lee, J. C. Carver, and A. Bosu, “Understanding the Impressions, Motivations, and Barriers of One Time Code Contributors to FLOSS Projects: A Survey,” in in Proce. IEEE/ACM 39th International Conference on Software Engineering, 2017, pp. 187–197.

[40] A. Ouni, R. G. Kula, and K. Inoue, “Search-based peer reviewers recommendation in modern code review,” in Proc. - IEEE International Conference on Software Maintenance and Evolution, 2017, pp. 367–377.

[41] P. Thongtanunam, S. McIntosh, A. E. Hassan, and H. Iida, “Review participation in modern code review,” Empir. Softw. Eng., vol. 22, no. 2, pp. 768–817, 2017.

[42] T. Baum, O. Liskin, K. Niklas, and K. Schneider, “A Faceted Classification Scheme for Change-Based Industrial Code Review Processes,” Proc. - 2016 IEEE Int. Conf. Softw. Qual. Reliab. Secur. QRS 2016, pp. 74–85, 2016.

[43] N. Kitagawa, H. Hata, A. Ihara, K. Kogiso, and K. Matsumoto, “Code Review Participation: Game Theoretical Modeling of Reviewers in Gerrit Datasets,” in Proc. 9th International Workshop on Cooperative and Human Aspects of Software Engineering, 2016, pp. 64–67.

[44] Y. Yu, H. Wang, G. Yin, and T. Wang, “Reviewer recommendation for pull-requests in GitHub: What can we learn from code review and bug assignment?,” Inf. Softw. Technol., vol. 000, pp. 1–15, 2015.

[45] S. McIntosh and Y. Kamei, “Are Fix-Inducing Changes a Moving Target? A Longitudinal Case Study of Just-In-Time Defect Prediction,” IEEE Trans. Softw. Eng., vol. 44, no. 5, pp. 412–428, 2018.

[46] R. Morales, S. McIntosh, and F. Khomh, “Do code review practices impact design quality? A case study of the Qt, VTK, and ITK projects,” in Proc. IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER), 2015, pp. 171–180.

[47] N. Fatima and K. Lumpur, “Individual , Social and Personnel Factors Influencing Modern Code Review Process.”

[48] O. Kononenko, T. Rose, O. Baysal, M. Godfrey, D. Theisen, and B. De Water, “Studying Pull Request Merges : A Case Study of Shopify ’ s Active Merchant,” in Proc. 40th International Conference on Software Engineering: Software Engineering in Practice, 2018, pp. 124–133.

[49] H. Lal and G. Pahwa, “Code review analysis of software system using machine learning techniques,” in in Proc. 11th International Conference on Intelligent Systems and Control, 2017, pp. 8–13.

[50] X. Yang, N. Yoshida, R. Gaikovina Kula, and H. Iida, “Peer Review Social Network (PeRSoN) in open source projects,” IEICE Trans. Inf. Syst., vol. E99D, no. 3, pp. 661–670, 2016.

[51] A. J. A. M. Van Deursen, C. Verlage, and E. Van Laar, “Social Network Site Skills for Communication Professionals: Conceptualization, Operationalization, and an Empirical Investigation,” IEEE Trans. Prof. Commun., vol. 62, no. 1, pp. 43–54, 2019.

[52] T. Baum, O. Liskin, K. Niklas, and K. Schneider, “Factors influencing code review processes in industry,” Proc. 2016 24th ACM SIGSOFT Int. Symp. Found. Softw. Eng. - FSE 2016, pp. 85–96, 2016.

[53] T. Zhang, M. Song, J. Pinedo, and M. Kim, “Interactive code review for systematic changes,” Proc. - Int. Conf. Softw. Eng., vol. 1, pp. 111–122, 2015.

[54] Z. Xia, H. Sun, J. Jiang, X. Wang, and X. Liu, “A Hybrid Approach to Code Reviewer Recommendation with Collaborative Filtering,” in SoftwareMining 2017, Urbana-Champaign, IL, USA, 2017, pp. 24–31.

[55] F. Armstrong, F. Khomh, and B. Adams, “Broadcast vs. Unicast Review Technology: Does It Matter?,” in in Proc. 10th IEEE International Conference on Software Testing, Verification and Validation, 2017, pp. 219–229.

[56] M. Beller, A. Bacchelli, A. Zaidman, and E. Juergens, “Modern code reviews in open-source projects: which problems do they fix?,” in Proc. 11th Working Conference on Mining Software Repositories, 2014, pp. 202–211.

[57] J. Jiang, Y. Yang, J. He, X. Blanc, and L. Zhang, “Who should comment on this pull request? Analyzing attributes for more accurate commenter recommendation in pull-based development,” Inf. Softw. Technol., vol. 84, pp. 48–62, 2017.

[58] D. Spadini, A. Bacchelli, M. Bruntink, F. Palomba, and L. Pascarella, “Information Needs in Contemporary Code Review,” in Proc. ACM on Human-Computer Interaction, 2018, vol. 2, no. CSCW, pp. 1–27.

[59] R. Wen, D. Gilbert, M. G. Roche, and S. McIntosh, “BLIMP tracer: Integrating build impact analysis with code review,” in Proc. IEEE International Conference on Software Maintenance and Evolution, 2018, pp. 685–694.

[60] Y. Yu, H. Wang, V. Filkov, P. Devanbu, and B. Vasilescu, “Wait for It: Determinants of pull request evaluation latency on GitHub,” in Proc. IEEE International Working Conference on Mining Software Repositories, 2015, vol. 2015-Augus, pp. 367–371.

[61] A. Bosu, “Modeling Modern Code Review Practices in Open Source Software Development Organizations,” in Proc. IDoESE ’13 Baltimore, Maryland USA, 2013.

[62] L. Claytor and F. Servant, “Understanding and Leveraging Developer Inexpertise,” Proc. 40th Int. Conf. Softw. Eng. Companion Proceeedings, pp. 404–405, 2018.

[63] F. Ebert, F. Castor, N. Novielli, and A. Serebrenik, “Confusion in Code Reviews: Reasons, Impacts, and Coping Strategies,” SANER 2019 -

Proc. 2019 IEEE 26th Int. Conf. Softw. Anal. Evol. Reengineering, pp. 49–60, 2019.

[64] A. M. Peter C. Rigby, Yue Cai Zhu, Samuel M. Donadelli, “Quantifying and Mitigating Turnover-Induced Knowledge Loss: Case Studies of Chrome and a Project at Avaya,” in Proc. IEEE/ACM 38th International Conference on Software Engineering, 2016, pp. 1006–1016.

[65] S. Fakhoury, “The Effect of Poor Source Code Lexicon and Readability on Developers´s Cognitive Load,” in Proc. ICPC, 2018, pp. 286–296.

[66] P. Thongtanunam, S. McIntosh, A. E. Hassan, and H. Iida, “Revisiting code ownership and its relationship with software quality in the scope of modern code review,” Proc. 38th Int. Conf. Softw. Eng. - ICSE ’16, no. 1, pp. 1039–1050, 2016.

[67] C. Parnin et al., “The Top 10 Adages in Continuous Deployment,” IEEE Softw., vol. 34, no. 3, pp. 86–95, 2017.

[68] B. Floyd, T. Santander, and W. Weimer, “Decoding the Representation of Code in the Brain: An fMRI Study of Code Review and Expertise,” in Proc.-IEEE/ACM 39th International Conference on Software Engineering, 2017, pp. 175–186.

[69] A. Murgia et al., “The emotional side of software developers in JIRA,” pp. 480–483, 2016.

[70] J. Shimagaki, Y. Kamei, S. McIntosh, A. E. Hassan, and N. Ubayashi, “A study of the quality-impacting practices of modern code review at Sony mobile,” in Proc. 38th International Conference on Software Engineering Companion, 2016, pp. 212–221.

[71] Z. X. Li, Y. Yu, G. Yin, T. Wang, and H. M. Wang, “What Are They Talking About? Analyzing Code Reviews in Pull-Based Development Model,” J. Comput. Sci. Technol., vol. 32, no. 6, pp. 1060–1075, 2017.

[72] A. Bosu and J. C. Carver, “Impact of developer reputation on code review outcomes in OSS projects,” Proc. 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. pp. 1–10, 2014.

[73] A. Meneely et al., “An empirical investigation of socio-technical code review metrics and security vulnerabilities,” 6th Int. Work. Soc. Softw. Eng. SSE 2014 - Proc., pp. 37–44, 2014.

[74] Y. Jiang, B. Adams, and D. M. German, “Will my patch make it? And how fast?: Case study on the linux kernel,” in Proc. IEEE International Working Conference on Mining Software Repositories, 2013, no. Section II, pp. 101–110.

[75] J. Tsay, L. Dabbish, and J. Herbsleb, “Influence of Social and Technical Factors for Evaluating Contribution in GitHub,” in Proc. International Conference on Software Engineering, 2014, pp. 356–366.

[76] A. Luxton-reilly, A. Lewis, and B. Plimmer, “Comparing Sequential and Parallel Code Review Techniques for Formative Feedback,” in Proc. 20th Australasian Computing Education Conference, 2018, pp. 45–52.

[77] J. Kim and E. Lee, “Understanding review expertise of developers: A reviewer recommendation approach based on latent Dirichlet allocation,” Symmetry (Basel)., vol. 10, no. 4, pp. 5–7, 2018.

[78] M. M. Rahman and C. K. Roy, “Impact of continuous integration on code reviews,” IEEE Int. Work. Conf. Min. Softw. Repos., pp. 499–502, 2017.

[79] T. Baum, F. Kortum, K. Schneider, A. Brack, and J. Schauder, “Comparing pre-commit reviews and post-commit reviews using process simulation,” J. Softw. Evol. Process, vol. 29, no. 11, pp. 1–15, 2017.

Situational Factors for Modern Code Review to Support Software Engineers' Sustainability

Sumaira Nazir[1], Nargis Fatima[2], Suriayati Chuprat[3]
Razak Faculty of Technology and Informatics
University Technology Malaysia (UTM), Kuala Lumpur, Malaysia[1, 2, 3]
Department of Engineering and Computer Science
National University of Modern Languages (NUML), Islamabad, Pakistan[1, 2]

***Abstract*—Software engineers working in Modern Code Review (MCR) are confronted with the issue of lack of competency in the identification of situational factors. MCR is a software engineering activity for the identification and fixation of defects before the delivery of the software product. This issue can be a threat to the individual sustainability of software engineers and it can be addressed by situational awareness. Therefore, the objective of the study is to identify situational factors concerning the MCR process. Systematic Literature Review (SLR) has been used to identify situational factors. Data coding along with continuous comparison and memoing procedures of grounded theory and expert review has been used to produce an exclusive and validated list of situational factors grouped under categories. The study results conveyed 23 situational factors that are grouped into 5 broad categories i.e. People, Organization, Technology, Source Code and Project. The study is valuable for researchers to extend the research and for software engineers to identify situations and sustain for longer.**

***Keywords*—Situational; modern code review; sustainable software engineer**

I. Introduction

Sustainable software engineering is presently a major concern in software development [1], [2]. It has five major aspects such as individual, social, economic, environmental, and technical [3], [4]. It is argued that work has been done regarding technical, economic, social, and environmental aspect of sustainable software engineering, however, individual sustainability aspect has been given less attention by the researchers and it warrants future research [2], [3], [4], [5], [6], [7].

Regarding individual sustainability, the software engineers are confronting with the issue of lack of competency in the identification of situational factors in various software engineering activities such as software requirement gathering and design, software construction and testing, modern code review [3], [8], [9].

The existing work concerning identification of situational factors have got attention by the researcher in software requirement and for software development, however, less attention has been devoted concerning situational context in modern code review specifically, to support software engineers' sustainability [3], [8], [9], [10], [11], [12] that can be the reason of software failure [3], [11], [13], [14].

Therefore, there is a need to identify situational factors for Modern Code Review (MCR) to overcome the issue of lack of competency in the identification of situational factors and to ensure the software engineers' sustainability involved in the MCR process [8], [9], [15]. MCR is an enhancement of Fagan's inspection, commonly known as a lightweight review process [16], [17]. It is supposed as a significant tool for improving code and software quality [16], [17], [18]. In MCR, software engineers i.e. authors and reviewers both aimed to improve the code quality [16], [17]. In this process, the code is reviewed by the reviewer from varying aspects. For instance, code style, code logic, code complexity etc. [16], [17], [18]. The process is highly reliant on review tools such as Code flow, Review board, Gerrit, etc. [16], [19].

This research has twofold aims i.e. to perform Systematic literature review (SLR) to identify situational factors for the MCR process and to validate them through expert review. The study detailed the SLR phases and expert review to identify and validate the situational factors for the MCR process that can help software engineers to sustain longer. This inquiry at one hand allows the investigators to outspread the research and on the other hand, it supports software engineers' sustainability through situational awareness in MCR.

The paper is arranged as Section II provides the background details. Section III details the study methodology. Section IV presents the results. Section V delivers the discussion. Section VI provides the conclusion and future work. Section VII deliberates the research contributions.

II. Background

Sustainability in software engineering is a noteworthy part of practices in the disciplines [1], [2]. It is defined as the "capacity to endure" [4]. There are five sustainability aspects reported in the literature such as individual, social, economic, environmental and technical [3], [4], [13], [14].

Economic sustainability aspects deals with investments and profitability [3], [4], [20], [21]. The technical sustainability aspect is connected to the ability to maintain and evolve software [8], [22], [23]. The social sustainability aspect is associated with the relationship between organizations, groups, and individuals [3], [4]. The environmental sustainability aspect is related to the objective to lessen the negative influence of software engineering activities on the environment [2], [3], [24], [25]. The individual sustainability aspect is related to well-being, education, and liberty of software engineers to sustain for longer [3], [4], [14]. Although valued work has been performed concerning to sustainable software engineering aspects i.e. technical, economic, social, and environmental, however, individual sustainability aspect has got less attention by the researchers and needs a detailed insight [2], [3], [4], [5], [6], [7].

Presently, software engineers are confronted with the issue concerning their sustainability such as lack of competency in the identification of situational factors in the Modern Code Review (MCR) process [3], [8], [9], [13], [14]. MCR is an important software engineering activity also known as a lightweight version of Fagan's inspection where developer other than source code author review the source code and guide the author in improving the quality of source code [16], [17], [26], [27], [28]. The process is performed with the help of review tools, for instance, Gerrit [16]. The overview of the process is shown in Fig. 1.

One of the reasons behind the issue of lack of competency in the identification of situational factors is unfamiliar situations [9], [23]. It is stated that the above issue can cause a decrease in the competency and capability of software engineers towards problem understanding, and identification of unfamiliar situations [9], [15]. This issue can be addressed by situational awareness in MCR [9], [11], [14].

In modern software development, situation-aware computing is extremely desirable [9]. Situation aware software engineering also called situational software engineering allows the software engineers to be able to deal with familiar situations instead of being unproductive with unfamiliar situations [9], [10], [11], [15], [29]. Situational software engineering ensures the software engineers' sustainability [9], [11], [15].

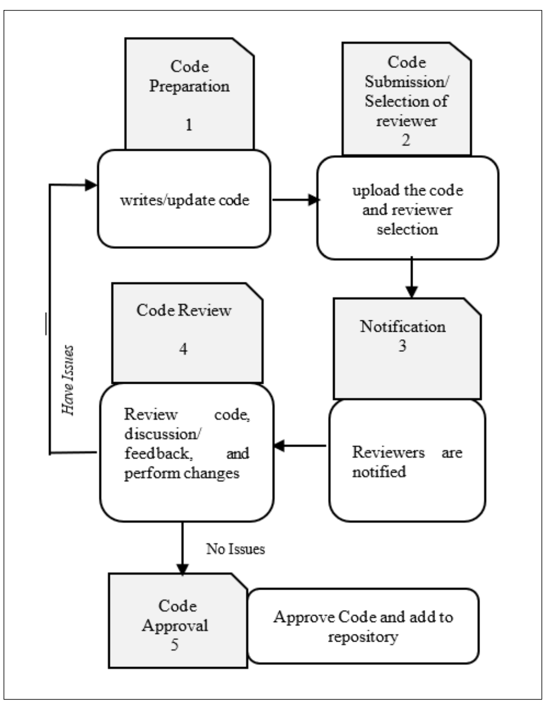


Fig. 1. MCR Overview [17].

The previous work established that researchers have highlighted the significance of situational factors identification and attention has been given on the identification of situational factors in software requirement and for software development [3], [8], [9], [10], [11], [12], however, little indication is available concerning to Modern Code Review (MCR) [30], [31], [32], [33], [34]. It results in the unavailability of situational guidelines that can help software engineers to increase their competency for the identification of situational factors. Therefore, this study aims to identify the unique and validated list of situational factors for the MCR process to support software engineers' sustainability.

III. Research Methodology

The research activities completed to produce the unique and validated list of situational factors for the MCR process to support software engineers' sustainability are explained in subsections.

A. Systematic Literature Review (SLR)

The Systematic Literature Review (SLR) technique specified by [35] has been employed to recognize the situational factors for the MCR process to support software engineers' sustainability. The SLR technique is a schematized approach to accomplish impartial results [35]. It involves significant steps such as SLR planning, SLR execution and, reporting the SLR results. It is a suitable technique to record significant data from existing research. The steps involved in SLR are explained in subsections.

1) Research questions: Designing the research question is an important aspect of SLR. For this study, the PICOC strategy proposed by [36] has been utilized to prepare the research question. PICOC stands for population, intervention, comparison, outcome, and context. As this study is inclusive of any type of comparison between methods, therefore the study selected only population, intervention, outcome, and context that is PIOC. Table I presents a summary of the PIOC strategy.

To collect the indications on the current state of research concerning situational factors, the planned research question is given below.

RQ1: What situational factors of the MCR process should be known, to support the software engineers' sustainability?

2) Search technique: The search technique includes the recognition of main search terms, finding their alternative and then constructing the search thread to search relevant data from data sources. Table II presents the main terms and their alternatives.

The search thread is planned based on the main terms and their alternate terms. The planned search thread is given below.

('conditional', 'contextual', 'situational factor') AND ('modern code inspection', 'contemporary code review' 'code review', 'code inspection', ' 'lightweight code inspection') AND ('sustainable software engineers', 'sustainable software developer', 'sustainable software programmer')

TABLE. I. POIC SUMMARY

Population	Intervention	Outcome	Context
Software Engineers	MCR Process	Situational factors for MCR process to support software engineers' sustainability	The research includes all types of studies, for instance, interviews, surveys, questionnaires, experiments, and case studies regarding MCR.

TABLE. II. MAIN TERMS AND THEIR ALTERNATES

Situational	Modern Code Review	Sustainable Software Engineer
'conditional', 'contextual', 'situational factor'	'modern code inspection', 'contemporary code review' 'code review', 'code inspection', 'lightweight code inspection'	'sustainable software developer', 'sustainable software programmer'

3) Data source: The data is collected from varying sources known for publishing software engineering research articles. The data origin utilized for the study includes ACM, IEEE, Springer link, Science direct, Wiley online, Scopus, and Web of Science. Papers published in 2013 to 2019 are considered for selection. The journals' articles, workshop papers, conference papers, book chapters, published thesis, and technical reports are searched in the defined databases. The data sources that reflect situational factors that impact the sustainability of software engineers involved in the MCR process are recognized as probably pertinent.

4) Study inclusion and exclusion principles: The inclusion criteria for including the relevant studies is as follows.

a) Research published in journals, conference proceedings, workshops, book chapters, thesis, or technical reports that are discussing situational factors for the MCR process to support software engineers.

b) Publication content is available completely.

c) Publications from 2013 to 2019.

d) Research papers are written in the English language.

The research papers were excluded based on the exclusion criteria specified below.

a) Research papers giving information such as conference proceedings, workshops table of content, and irrelevant title.

b) Research papers that do not contain any one of the study main terms or their alternates terms.

c) Duplicate research papers

Fig. 2 summarizes the study inclusion and exclusion principles.

5) Quality assessment: The selected research papers are weighed for their quality by using the checklist provided by [35]. The checklist used for evaluating the quality of the research papers is given in Table III. Furthermore, each question given in the checklist presented in Table III is answered by the measures given by [35]. The measures are shown in Table IV.

6) Data extraction: The data is extracted from the selected studies with the help of extraction forms given by [35]. The details about the data extraction form are presented in Table V.

TABLE. III. QUALITY ASSESSMENT CHECKLIST [36]

Question	Answer
Are the objective clearly stated?	Yes/ No/Partially
Are the findings sound and significant?	-
Are the prediction techniques used clearly described and their selection are justified?	-
Is the facts been extended through the research?	-
Is the multiplicity of viewpoint and background been explored?	-
Are the associations between data, interpretation, and conclusions are vibrant?	-
Does the depth of the data is conveyed?	-

TABLE. IV. MEASURES FOR ANSWERING QUESTION GIVEN IN CHECKLIST [36]

Answer	Score
Yes	1
No	0
Partially	0.5

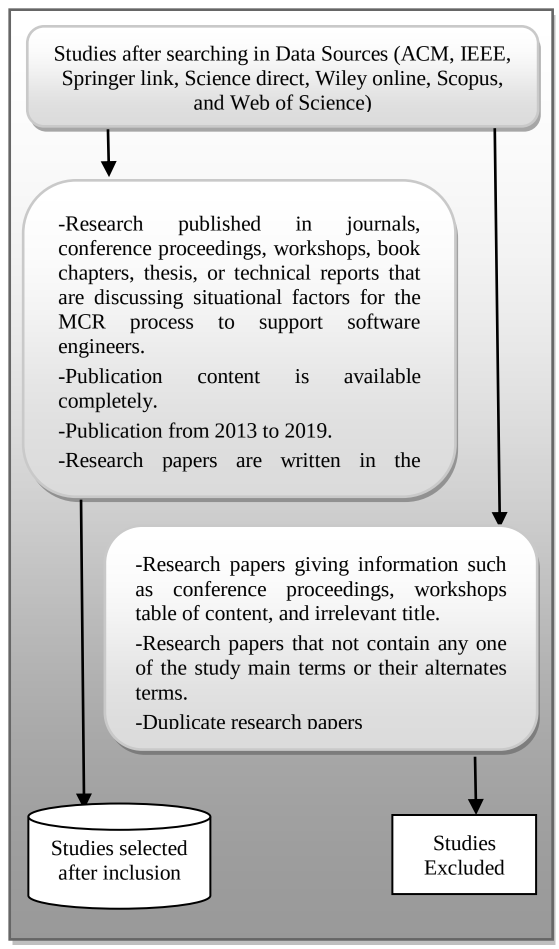


Fig. 2. Study Inclusion and Exclusion Criteria.

TABLE. V. DATA EXTRACTION FORM

Data Items	Data information	Notes
Paper ID	A unique identifier SFS<1--n>	
Title		
Author (s)		
Year		
Study type	(Conference/Journal/Book/Thesis)	
Study Publisher	IEEE	
Situational Factors		

7) Data analysis: After the data extraction, qualitative data analysis has been performed using the grounded theory given by [37], [38]. The grounded theory techniques i.e. data coding, continuous comparison and memoing has been used to recognize a unique list of situational factors grouped under various categories for the MCR process.

B. Expert Review

After finding the unique list of situational factors grouped under various categories, the list has been sent to the experts for the assessment concerning naming, terms, and classification of identified situational factors. The experts are also asked to mention new situational factors. The experts are designated based on their expertise in MCR along with software development experience for more than 10 years, the familiarity of situational software engineering, sustainable software engineering, and individual sustainability. The guidelines given by Ayyub [39], [40] are followed. The final list of situational factors along with their classification is presented in Section VI. Fig. 3 highlights the summarized view of the methodology employed for the identification of situational factors for the MCR process to support software engineers' sustainability.

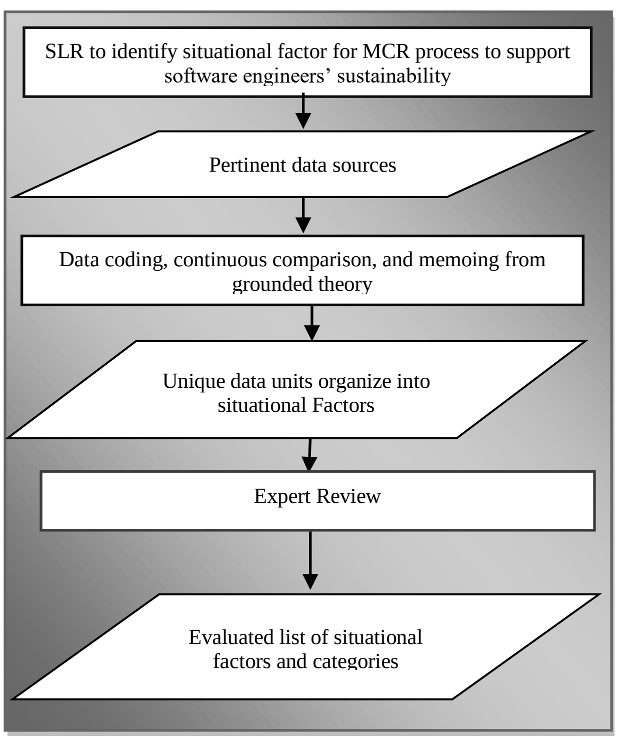


Fig. 3. Overview of Study Methodology.

IV. RESULTS

This section presents the results of SLR and the expert review i.e. the study selection process, distribution of data sources and an evaluated list of situational factors for the MCR process to support software engineers' sustainability.

A. Study Selection Process Results

In the initial search, 9295 papers are found based on defined study main terms. The studies mentioning exclusive information concerning table of content, workshop or conference preceding details or having disparate titles are eliminated, and 1096 studies are selected. The 1096 studies are assessed for relevancy concerning main terms of the study i.e. Situational, Modern Code Review, Sustainable Software Engineer. The studies not representing any of the study main terms are rejected and 187 studies are included. Afterward, 187 obtained studies are checked for replication. After replication assessment 162 studies are obtained for their quality assessment. After quality assessment 158 studies are included for deep review for the identification of situational factors for the MCR process to support software engineers' sustainability.

B. Distribution of Data Sources

Total 9295 papers obtained after an initial search from defined databases. Finally, 158 papers are selected after going through inclusion/exclusion and quality assessment.

C. Situational Factors

The study results reported 23 situational factors grouped into five categories i.e. people, organization, project, source code, and technology. The classification of the identified situational factors is discussed in subsections. Table VI summarizes the situational factors along with their classification for the MCR process to support software engineers' sustainability.

1) People: This category includes factors that are directly related to people. The situational factors included in this category are team, team interaction, reviewer response, and knowledge sharing [17], [19], [26], [34], [41], [42], [43], [44], [45], [46], [47], [48].

2) Source code: It refers to a list of human-readable instructions that a programmer writes using code editors. The source code runs through a compiler to turn it into machine code, that a computer can understand and execute [49]. The situational factors grouped under this category are source code attributes, source code change attributes, source code change documentation, testing, review concentration, and defect [17], [18], [19], [32], [43] [50], [51], [52], [53], [54], [55].

3) Organization: It refers to the group of people and facilities with an arrangement of tasks, authorities, and relations [56]. The situational factors included in this category are resources [26], organization policy, organization practices, organization standards, organization attributes, and information dissemination [26], [41], [43], [50], [57], [58].

4) Project: The project is an arrangement of tasks that are prearranged from beginning to end bounded by resources and required outcomes [56]. This category involves two situational factors i.e. project attributes and project release management [16], [17], [43], [59], [60].

TABLE. VI. SITUATIONAL FACTORS AND THEIR CATEGORIES

Categories	**Situational Factors**	**References**
People	Team	[17], [19], [26], [34], [41], [42], [43], [65], [66], [61], [67], [50], [68]
	Team Interaction	[17], [26], [42], [43], [65], [50], [69], [70],[59], [71], [72]
	Reviewer Response	[16], [17], [18], [26], [32], [42], [43], [66], [44], [45], [73]
	Knowledge Sharing	[26], [43], [61], [67], [69], [46], [47]
Source Code	Source Code Attributes	[17], [18], [19], [30], [31], [32], [33], [41] [42], [43], [66], [70], [59], [44], [47] [74], [51] , [52], [75], [76], [77], [66]
	Source Code Change Attributes	[16], [17], [18], [19] [26], [30], [32], [43] [50], [70], [59], [71] [51], [52], [75], [78]
	Source Code Change Documentation	[16], [18], [19], [26], [30], [42], [43], [65] [50], [59], [51], [76], [79]
	Testing	[18], [19], [50], [51], [52], [53], [54]
	Review Concentration	[17], [32], [42], [55]
	Defect	[32], [41], [70], [71], [55], [62]
Organization	Resources	[26]
	Organization Policy	[26], [57]
	Organization Practices	[26], [41], [43], [50], [57], [58]
	Organization Standards	[17], [18], [26], [41], [61], [57]
	Organization Attributes	[17], [26], [41], [61]
	Information Dissemination	[57]
Project	Project Attributes	[16], [17], [43], [59]
	Project Release Management	[17], [18], [43],
Technology	Process	[17], [18], [26], [31], [34] [41] [43], [61], [71], [57]
	Tool	[18], [26] [34], [61], [50], [69], [70], [59], [71], [74], [62], [63], [64]
	Technology Maturity	[26]
	Technology Accessibility	[26], [61]
	Training	[17], [18], [26], [70]

5) Technology: It refers to the approaches, skills, and processes used in the creation of goods or services in the achievement of aims [49]. The situational factors included in this category are process, tool, technology maturity, technology accessibility, and training [18], [26], [31], [34] [41] [43], [61], [62], [63], [64].

V. DISCUSSION

This study has provided a comprehensive list of classified and validated situational factors for the MCR process to support software engineers' sustainability through SLR and expert review. The identified situational factors that can impact the sustainability of software engineers can be an important reference for researchers involved in research concerning situational software engineering, sustainable software engineering, and MCR. The work can support the sustainability of software engineers involved in the MCR process by providing the list of situational factors. The identified list can also act as a guide for the researchers and practitioners working in situational software engineering, and sustainable software engineering.

The study presents the situational factors based on literature. Although effort has been made to cover all the related research papers to present the comprehensive list of situational factors for MCR process to support software engineers' sustainability, however, there can be a possibility that some research may not be covered.

VI. CONCLUSION AND FUTURE WORK

This work has provided a unique, classified, and validated list of situational factors for the MCR process to support software engineers' sustainability. A total of 23 situational factors have been identified as a result of this work. The identified factors are broadly grouped under five categories i.e. People, Organization, Process, Source code, and Technology. These factors can support the sustainability of software engineers.

In the future, a more inclusive list will be shaped, the ongoing research objectives. In addition to this, a comprehensive and enhanced MCR process will be produced with situational factors. This work provided situational factors for MCR activity of software engineering that allows the investigators to extend this research by determining other situational factors in other software engineering activities.

VII. CONTRIBUTION

The examination contributed to the software engineering body of knowledge (SWEBOK), Situational software engineering and sustainable software engineering by highlighting the worth of situational factor identification for the sustainability of the software engineers involved in MCR.

REFERENCES

[1] B. Penzenstadler and H. Femmer, "Towards a Definition of Sustainability in and for Software Engineering at Sustainable System : Product," p. 2013, 2013.

[2] S. Naumann, E. Kern, M. Dick, and T. Johann, "Sustainable Software Engineering: Process and Quality Models, Life Cycle, and Social Aspects," ICT Innov. Sustain. Adv. Intell. Syst. Comput., vol. 310, pp. 191–205, 2015.

[3] R. Chitchyan, I. Groher, and J. Noppen, "Uncovering sustainability concerns in software product lines," J. Softw. Evol. Process, vol. 29, no. 2, pp. 1–20, 2017.

[4] B. Penzenstadler et al., “Software Engineering for Sustainability: Find the Leverage Points!,” IEEE Softw., vol. 35, no. 4, pp. 22–33, 2018.

[5] S. Naumann, M. Dick, E. Kern, and T. Johann, “The GREENSOFT Model: A reference model for green and sustainable software and its engineering,” Sustain. Comput. Informatics Syst., vol. 1, no. 4, pp. 294–304, 2011.

[6] A. D. Komeil Raisian, Jamaiah Yahaya, “Sustainable Software Development Life Cycle Process Model Based on Capability Maturity Model Integration : a Study in Malaysia,” J. Threoretical Appl. Inf. Technol., vol. 95, no. 21, pp. 5723–5734, 2017.

[7] B. Penzenstadler, A. Raturi, D. Richardson, C. Calero, H. Femmer, and X. Franch, “Systematic Mapping Study on Software Engineering for Sustainability (SE4S),” in Proc. 18th International Conference on Evaluation and Assessment in Software Engineering, 2014, pp. 1–14.

[8] R. Chitchyan, L. Duboc, C. Becker, S. Betz, B. Penzenstadler, and C. C. Venters, “Sustainability Design in Requirements Engineering : State of Practice,” in IEEE/ACM 38th IEEE International Conference on Software Engineering, 2016, pp. 533–542.

[9] A. A. Abbood and G. Sulong, “Segmentation and Enhancement of Fingerprint,” Springer Int. Publ., vol. 5, no. 3, 2018.

[10] P. Clarke, R. V. O. Connor, R. V. O. Connor, and B. Leavy, “A complexity theory viewpoint on the software development process and situational context,” no. May, 2016.

[11] H. H. Khan and M. N. Malik, “Software Standards and Software Failures: A Review with the Perspective of Varying Situational Contexts,” IEEE Access, vol. 5, pp. 17501–17513, 2017.

[12] P. Clarke and R. V. O’Connor, “The situational factors that affect the software development process: Towards a comprehensive reference framework,” Inf. Softw. Technol., vol. 54, no. 5, pp. 433–447, 2012.

[13] S.Nazir, N. Fatima, and S. Chuprat “Individual Sustainability Barriers and Mitigation Strategies : Systematic Literature Review Protocol.”

[14] S. Nazir, N. Fatima, S. Chuprat, H. Sarkan, N. F. Nilam, and N.A. Sajarif, “Sustainable Software Engineering:A Perspective of Individual Sustainability,” Int. J. Adv. Sci. Eng. Inf. Technol.

[15] G. Marks, R. V. O’Connor, and P. M. Clarke, “The impact of situational context on the software development process – A case study of a highly innovative start-up organization,” Commun. Comput. Inf. Sci., vol. 770, pp. 455–466, 2017.

[16] A. Bacchelli and C. Bird, “Expectations, outcomes, and challenges of modern code review,” in Proc. International Conference on Software Engineering, 2013, pp. 712–721.

[17] A. Bosu, J. C. Carver, C. Bird, J. Orbeck, and C. Chockley, “Process Aspects and Social Dynamics of Contemporary Code Review: Insights from Open Source Development and Industrial Practice at Microsoft,” IEEE Trans. Softw. Eng., vol. 43, no. 1, pp. 56–75, 2017.

[18] O. Kononenko, O. Baysal, and M. W. Godfrey, “Code Review Quality: How Developers See It,” in Proc. International Conference on Software Engineering, 2016, pp. 1028–1038.

[19] A. Ram, Achyudh ; Sawant, Anand; Castelluccio, Marco; Bacchelli, “What Makes a Code Change Easier to Review? An Empirical Investigation on Code Change Reviewability,” in Proc. ESEC/FSE, 2018.

[20] M. L. Gibson, C. C. Venters, M. Palacin-silva, and N. Seyff, “Mind the chasm: A UK fisheye lens view of sustainable software engineering,” 2017.

[21] R. Ahmad, F. Baharom, and A. Hussain, “Software Sustainability Development : Impactibility Characteristic Focuses on Social Approach,” in Proc. 6th International Conference on Computing and Informatics, 2017, no. 093, pp. 595–600.

[22] H. Koziolek, “Sustainability Evaluation of Software Architectures : A Systematic Review,” in Proc. QoSA+ISARCS, 2011, pp. 3–12.

[23] C. Becker et al., “Requirements: The key to sustainability,” IEEE Softw., vol. 33, no. 1, pp. 56–65, 2016.

[24] E. Kern et al., “Sustainable software products—Towards assessment criteria for resource and energy efficiency,” Futur. Gener. Comput. Syst., vol. 86, no. 3715, pp. 199–210, 2018.

[25] I. Manotas et al., “An empirical study of practitioners’ perspectives on green software engineering,” in IEEE/ACM 38th IEEE International Conference on Software Engineering, 2016, pp. 237–248.

[26] L. MacLeod, M. Greiler, M. A. Storey, C. Bird, and J. Czerwonka, “Code Reviewing in the Trenches: Challenges and Best Practices,” IEEE Softw., vol. 35, no. 4, pp. 34–42, 2018.

[27] N. Fatima, S. Chuprat, and S. Nazir, “Challenges and Benefits of Modern Code Review-Systematic Literature Review Protocol,” in Proc. International Conference on Smart Computing and Electronic Enterprise, 2018, pp. 1–5.

[28] S. Nazir, N. Fatima, and S. Chuprat, “Modern Code Review Benefits–Primary findings of a systematic literature review,” in The 3rd International Conference on Software Engineering and Information Management.

[29] C. K. Chang, “Situation Analytics: A Foundation for a New Software Engineering Paradigm,” Computer (Long. Beach. Calif)., vol. 49, no. 1, pp. 24–33, 2016.

[30] E. W. dos Santos and I. Nunes, “Investigating the Effectiveness of Peer Code Review in Distributed Software Development,” in Proc. 31st Brazilian Symposium on Software Engineering, 2017, pp. 84–93.

[31] T. Baum, F. Kortum, K. Schneider, A. Brack, and J. Schauder, “Comparing pre-commit reviews and post-commit reviews using process simulation,” J. Softw. Evol. Process, vol. 29, no. 11, pp. 1–15, 2017.

[32] F. Armstrong, F. Khomh, and B. Adams, “Broadcast vs. Unicast Review Technology: Does It Matter?,” in Proc. 10th IEEE International Conference on Software Testing, Verification and Validation, 2017, pp. 219–229.

[33] F. Ebert, F. Castor, N. Novielli, and A. Serebrenik, “Confusion in Code Reviews: Reasons, Impacts, and Coping Strategies,” SANER 2019 - Proc. 2019 IEEE 26th Int. Conf. Softw. Anal. Evol. Reengineering, pp. 49–60, 2019.

[34] C. Sadowski, E. Söderberg, L. Church, M. Sipko, and A. Bacchelli, “Modern code review: : A Case Study at Google,” in Proc. ACM/IEEE 40th International Conference on Software Engineering: Software Engineering in Practice, 2018, pp. 181–190.

[35] B. Kitchenham and S. Charters, “Source: " Guidelines for performing Systematic Literature Reviews in SE " , Kitchenham et al Guidelines for performing Systematic Literature Reviews in Software Engineering Source: " Guidelines for performing Systematic Literature Reviews i,” pp. 1–44, 2007.

[36] F. Terms, “ M. Petticrew and H. Roberts. Systematic Reviews in the Social Sciences: A Practical Guide . Oxford: Blackwell 2006. 352 pp. ISBN 1 4051 2110 6. £29.99 ,” Couns. Psychother. Res., vol. 6, no. 4, pp. 304–305, 2006.

[37] Kathy Charmaz, Constructing Grounded Theory, A practical Guide through Qualitative Analysis. 2007.

[38] K.-J. Stol, P. Ralph, and B. Fitzgerald, “Grounded theory in software engineering research,” no. October 2017, pp. 120–131, 2016.

[39] B. Ayyub, “A practical guide on conducting expert-opinion elicitation of probabilities and consequences for corps facilities,” Inst. Water Resour. Alexandria, VA, USA, no. January, 2001.

[40] R. Boring, D. Gertman, J. Joe, and J. Marble, “Simplified expert elicitation guideline for risk assessment of operating events,” … Natl. Lab. INL …, no. June, p. 65, 2005.

[41] T. Baum, O. Liskin, K. Niklas, and K. Schneider, “Factors influencing code review processes in industry,” Proc. 2016 24th ACM SIGSOFT Int. Symp. Found. Softw. Eng. - FSE 2016, pp. 85–96, 2016.

[42] M. Di Biase, M. Bruntink, and A. Bacchelli, “A security perspective on code review: The case of chromium,” in Proc. IEEE 16th International Working Conference on Source Code Analysis and Manipulation, 2016, pp. 21–30.

[43] O. Kononenko, T. Rose, O. Baysal, M. Godfrey, D. Theisen, and B. De Water, “Studying Pull Request Merges : A Case Study of Shopify ’ s Active Merchant,” in Proc. 40th International Conference on Software Engineering: Software Engineering in Practice, 2018, pp. 124–133.

[44] S. McIntosh and Y. Kamei, “Are Fix-Inducing Changes a Moving Target? A Longitudinal Case Study of Just-In-Time Defect Prediction,” IEEE Trans. Softw. Eng., vol. 44, no. 5, pp. 412–428, 2018.

[45] S. Ruangwan, P. Thongtanunam, A. Ihara, and K. Matsumoto, “The Impact of Human Factors on the Participation Decision of Reviewers in Modern Code Review,” Empir. Softw. Eng. Manuscr., pp. 1–43, 2018.

[46] C. Parnin et al., “The Top 10 Adages in Continuous Deployment,” IEEE Softw., vol. 34, no. 3, pp. 86–95, 2017.

[47] X. Yang, R. G. Kula, N. Yoshida, and H. Iida, “Mining Code Review Repositories : People , Process and Product,” IEEE Work. Conf. Min. Softw. Repos., pp. 16–19, 2016.

[48] S. Nazir, N. Fatima, and S. Chuprat, “Situational factors affecting Software Engineers Sustainability: A Vision of Modern Code Review,” in 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) , in press.

[49] ISO/IEC and IEEE, “ISO/IEC/IEEE 24765:2010 - Systems and software engineering -- Vocabulary,” Iso/Iec Ieee, vol. 2010, p. 410, 2010.

[50] G. Gousios, M.-A. Storey, and A. Bacchelli, “Work practices and challenges in pull-based development,” in Proc. 38th International Conference on Software Engineering, 2016, pp. 285–296.

[51] S. Fakhoury, “The Effect of Poor Source Code Lexicon and Readability on Developers´s Cognitive Load,” in Proc. ICPC, 2018, pp. 286–296.

[52] D. Spadini, A. Bacchelli, M. Bruntink, F. Palomba, and L. Pascarella, “Information Needs in Contemporary Code Review,” in Proc. ACM on Human-Computer Interaction, 2018, vol. 2, no. CSCW, pp. 1–27.

[53] J. Tsay, L. Dabbish, and J. Herbsleb, “Let’s talk about it: evaluating contributions through discussion in GitHub,” in Proc. 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering - FSE 2014, 2014, pp. 144–154.

[54] Y. Yu, H. Wang, V. Filkov, P. Devanbu, and B. Vasilescu, “Wait for It: Determinants of pull request evaluation latency on GitHub,” in Proc. IEEE International Working Conference on Mining Software Repositories, 2015, vol. 2015-Augus, pp. 367–371.

[55] C. Thompson and D. Wagner, “A Large-Scale Study of Modern Code Review and Security in Open Source Projects,” Proc. 13th Int. Conf. Predict. Model. Data Anal. Softw. Eng., pp. 83–92, 2017.

[56] I. Standard, ISO/IEC/IEEE Systems and software engineering -- System life cycle processes, vol. 8. 2013.

[57] D. M. German, U. Rey, and J. Carlos, “‘ Was my contribution fairly reviewed ?’ A Framework to Study the Perception of Fairness in Modern Code Reviews,” in Proc. ACM/IEEE 40th International Conference on Software Engineering Synthesizing, 2018, no. 2, pp. 523–534.

[58] M. Barnett, C. Bird, J. Brunet, and S. K. Lahiri, “Helping developers help themselves: Automatic decomposition of code review changesets,” Proc. - Int. Conf. Softw. Eng., vol. 1, no. August 2014, pp. 134–144, 2015.

[59] P. C. Rigby and C. Bird, “Convergent Contemporary Software Peer Review Practices Categories and Subject Descriptors,” in Proc. ESEC/FSE, 2013, pp. 202–212.

[60] S. Nazir, N. Fatima, and S. Chuprat, “Does Project Associated Situational Factors have Impact on Sustainability of Modern Code Review Workforce?,” in 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) , in press.

[61] A. Kalyan, M. Chiam, J. Sun, and S. Manoharan, “A Collaborative Code Review Platform for GitHub,” in Proc. IEEE International Conference on Engineering of Complex Computer Systems, 2017, pp. 191–196.

[62] D. Singh, V. R. Sekar, K. T. Stolee, and B. Johnson, “Evaluating how static analysis tools can reduce code review effort,” IEEE Symp. Vis. Lang. Human-Centric Comput., pp. 101–105, 2017.

[63] Z. X. Li, Y. Yu, G. Yin, T. Wang, and H. M. Wang, “What Are They Talking About? Analyzing Code Reviews in Pull-Based Development Model,” J. Comput. Sci. Technol., vol. 32, no. 6, pp. 1060–1075, 2017.

[64] Y. Yu, H. Wang, G. Yin, and T. Wang, “Reviewer recommendation for pull-requests in GitHub: What can we learn from code review and bug assignment?,” Inf. Softw. Technol., vol. 000, pp. 1–15, 2015.

[65] A. Bosu, J. Carver, R. Guadagno, B. Bassett, D. McCallum, and L. Hochstein, “Peer impressions in open source organizations: A survey,” J. Syst. Softw., vol. 94, pp. 4–15, 2014.

[66] A. Ouni, R. G. Kula, and K. Inoue, “Search-based peer reviewers recommendation in modern code review,” in Proc. - IEEE International Conference on Software Maintenance and Evolution, 2017, pp. 367–377.

[67] N. Kitagawa, H. Hata, A. Ihara, K. Kogiso, and K. Matsumoto, “Code Review Participation: Game Theoretical Modeling of Reviewers in Gerrit Datasets,” in Proc. 9th International Workshop on Cooperative and Human Aspects of Software Engineering, 2016, pp. 64–67.

[68] C. Bird, T. Carnahan, and M. Greiler, “Lessons learned from building and deploying a code review analytics platform,” IEEE Int. Work. Conf. Min. Softw. Repos., pp. 191–201, 2015.

[69] T. Baum, O. Liskin, K. Niklas, and K. Schneider, “A Faceted Classification Scheme for Change-Based Industrial Code Review Processes,” Proc. - 2016 IEEE Int. Conf. Softw. Qual. Reliab. Secur. QRS 2016, pp. 74–85, 2016.

[70] F. Ebert, F. Castor, N. Novielli, and A. Serebrenik, “Confusion detection in code reviews,” in Proc. IEEE International Conference on Software Maintenance and Evolution, 2017, pp. 549–553.

[71] H. Lal and G. Pahwa, “Code review analysis of software system using machine learning techniques,” in in Proc. 11th International Conference on Intelligent Systems and Control, 2017, pp. 8–13.

[72] F. Ebert, F. Castor, N. Novielli, and A. Serebrenik, “Communicative Intention in Code Review Questions,” 2018.

[73] P. Thongtanunam, S. McIntosh, A. E. Hassan, and H. Iida, “Investigating code review practices in defective files: An empirical study of the Qt system,” IEEE Int. Work. Conf. Min. Softw. Repos., vol. 2015-Augus, pp. 168–179, 2015.

[74] J. Kim and E. Lee, “Understanding review expertise of developers: A reviewer recommendation approach based on latent Dirichlet allocation,” Symmetry (Basel)., vol. 10, no. 4, pp. 5–7, 2018.

[75] A. Luxton-reilly, A. Lewis, and B. Plimmer, “Comparing Sequential and Parallel Code Review Techniques for Formative Feedback,” in Proc. 20th Australasian Computing Education Conference, 2018, pp. 45–52.

[76] A. Bosu, M. Greiler, and C. Bird, “Characteristics of useful code reviews: An empirical study at Microsoft,” in Proc. IEEE International Working Conference on Mining Software Repositories, 2015, vol. 2015-Augus, pp. 146–156.

[77] Z. Xia, H. Sun, J. Jiang, X. Wang, and X. Liu, “A Hybrid Approach to Code Reviewer Recommendation with Collaborative Filtering,” in SoftwareMining 2017, Urbana-Champaign, IL, USA, 2017, pp. 24–31.

[78] M. Beller, A. Bacchelli, A. Zaidman, and E. Juergens, “Modern code reviews in open-source projects: which problems do they fix?,” in Proc. 11th Working Conference on Mining Software Repositories, 2014, pp. 202–211.

[79] O. Kononenko, O. Baysal, L. Guerrouj, Y. Cao, and M. W. Godfrey, “Investigating code review quality: Do people and participation matter?,” in Proc. IEEE 31st International Conference on Software Maintenance and Evolution, 2015, pp. 111–120.

Plant Disease Detection using Internet of Thing (IoT)

Muhammad Amir Nawaz[1], Tehmina khan[2], Rana Mudassar Rasool[3], Maryam Kausar[4]
Amir Usman[5], Tanvir Fatima Naik Bukht[6], Rizwan Ahmad[7], Jaleel Ahmad[8]
Department of Computer Science, Institute of Southern Punjab Multan, Pakistan[1,2,3,4,6]
Department of Computer Science, NCBA&E Multan, Pakistan[5]
Department of Computer Science, Ripah International University Lahore, Pakistan[7,8]

***Abstract*—This paper presents the idea of internet of things (IOT) innovation to percept data, and talks about the job of the IOT innovation in farming infection and bug nuisance control, which incorporates rural ailment and bug checking framework, gathering sickness and creepy crawly bother data utilizing sensor hubs, information preparing and mining, etc. A malady and bug irritation control framework dependent on IOT is proposed, which comprised of three levels and three frameworks. The framework can give another approach to get to horticultural data for the farm. In this paper a computerized framework has been created to decide if the plant is ordinary or infected. The typical development of the plants, yield and nature of horticultural items is truly influenced by plant illness. This paper attempt to build up a robotized framework that identifies the nearness of disease in the plants. A mechanized ailment recognition framework is created utilizing sensors like temperature, moistness and shading dependent on variety in plant leaf wellbeing condition. The qualities dependent on temperature, mugginess and shading parameters are utilized to distinguish nearness of plant sickness.**

Keywords*—*Plant diseases; internet of things; temperature sensor plants; farming

I. Introduction

A. Background and Motivation

Pakistan is a place that is known for agribusiness. Two-third of populace depends upon farming for their work. It is the fundamental establishment of financial improvement of the nation. The agribusiness likewise gives work chances to exceptionally huge level of populace.

Plant health condition assumes a crucial job to acquire great benefit for the ranchers. Legitimate checking of plant health is required at various phases of plant development so as to anticipate malady influencing plants. Presence of vermin and malady influence the estimation of harvest development and limits yield significantly. Present day framework relies upon unaided eye perception which is a tedious procedure. Programmed recognition of plant infection can be embraced to distinguish plant illness at beginning times. Different illness the board methodologies have been utilized by farmer at customary intervals so as to avoid plant diseases. Research in agriculture robots has been developing in the most recent years, on account of potential applications and industry endeavors in robot improvement. Their job was explored for some agrarian errands, chiefly engaged in expanding mechanization of traditional farming machines and making procedures, for example, progress planning, seeding, preparation, and reaping. Efficient, dreary, and time-subordinate errands appear to speak to the best fields of use for robots, particularly in an arable cultivating setting with brief harvests. Close to agronomic practices, automated plant assurance has likewise been explored, yet may speak to the most intricate test for analysts and engineers since inquiries in respect to pathogen determination must be considered alongside normal robot-related issues. As of late, look into in programmed acknowledgment of sicknesses has been quickly developing, with potential applications for creating robots ready to perceive single plants, find and distinguish ailments, and begin schedules for ailment the board. This paper expects to give subtleties of that new age of robots that could bolster plant pathologists.

B. Internet of Things (IoT) in Plants

The overall blueprint of Internet of things helper parts showed of mind blowing ability in those headways in spaces of plants and the continuous example of Accuracy in plants. Continuous improvements in sensor development, nearby downsizing in equipment in addition remarkable descent in their rate have donated the ton to a mechanical improvement of customary cultivating to correctness besides little scale exactness plants. Air sensor, ground sensor, radiation sensors, atmosphere station highlight that it is almost sensor as well as sensors data streams, which are secured as well as use for checking, data excavating, thinking, and control. Besides, starting late, there is a growing enthusiasm for high gauge as well as harmless country things. That example has produced the necessity for cover operable, spread, incredible, and careful co-appointments tractability systems. The IoT group of advances gives all the suitable instruments to building and keeping up such foundation and administrations, extraordinarily intended to help supply chains in rural and plant sectors Sensors, as wired and remote sensors, have been broadly utilized in horticulture during the most recent decades.

Detecting the earth in which creation happens, and, all the more as of late, the reactions of the plants to the atmosphere is urgent for taking the right and increasingly exact choices, streamlining profitability and nature of the cultivars. Adaptable gadgets, with high computational capacities, exceptionally advantageous structure factor and ease, would nowadays be able to be utilized, on batteries, and work for significant lots, with or without the help of intensity gathering modules. What's more, present day implanted gadgets have adequate assets to bolster all the more requesting sensors, for example, picture sensors, and the help of progressively modern systems administration conventions, such TCP/IP, broadening the customary systems administration abilities.

Web of Stuffs is quickly progressing then various innovative presentations as well as organizations remain ascending out of that. The ton of exploration is actuality coordinating in the direction of joining a various dissimilar arrangements, Security insistence by numerous degrees of Internet of things as well as examination that will provide an unrivaled comprehension into the "Colossal Data" in order to improve distinctive business structures. National course of action of governments around the world for extended creation pace of fresh cut vegetables and meat, at lower cost, with higher quality benchmarks, similarly as, the buyers' enthusiasm for straightforwardness in the age cycle and the biological impression of the things they buy, give IoT a tremendous field to progress and scattering. As shown by Bradley the examinations from 2015-2020 of potential IoT worth vary basically, running from at any rate $1 trillion up to more than $15 trillion, barring the extended salaries, the benefits of cost decline among associations and adventures and the general budgetary development due to IoT. A huge piece of the extra estimation of IoT starts from the versatility and the improvement and precision that it brings into the age methodology of industry and creation units of various sorts. Subsequently, it isn't so perilous to check that cultivating zone shapes at all levels will profoundly change as soon as possible. Concerning agribusiness, IoT is required to streamline the age by various strategies. Farmlands and nurseries are going to move from precision to a littler scale exactness model of country creation. Disseminated, inescapable registering and exact observing of the offices will give the ideal developing or living conditions for the two vegetables and creatures. Self-ruling frameworks will be capable not exclusively to direction the actuators in the most proficient manner, improving the utility and asset utilization, yet additionally to control the generation in understanding for a market circumstance, boosting the benefit as well as limiting expenses each technique under sun. Then again, sustenance supply chains, furnished with RFID gear, will most likely screen each phase in the life of an item, make programmed thinking, in the event of a broken item and increment customer's sentiment of wellbeing, through a straightforward item life cycle data framework.

Clearly the financial numbers identified with IoT are huge, enticing some intense players to put resources into it. Models, corresponding the progressing acquisition for Nest Laboratories, an association increasing pragmatic involvement of IoT in home computerization, through Google is dollar 4.02 billion in genuine money and the acquisition for Jasper Machineries, architect of as well as IoT will arrange, through Cisco with dollar 2.04 billion, reveal staggering ability for IoT as well as exhibit which is exceedingly speaking to tremendous money related experts and behemoth mechanical firms. The association course of action, regardless, isn't so irrelevant. This is because of the way that the organizations associated with IoT put resources into one or a couple of parts of it, due to its wide nature. Accordingly, at some point or another, they should collaborate with one another, setting aside any challenge, or the idea of who is increasingly significant, so as to present some widespread models in the advancing IoT publicity.

C. IoT in Plant Disease Detection

Identification of diseases in the plant is most extreme requirement for farmers and agriculture specialists. The principle point of the proposed framework is to recognize plant diseases with the help of IoT (Internet of Things). In the greater part of the plants the sickness beginning happens on plant leaves. Subsequently, in the proposed work we have considered location of plant sickness present on leaves. The segregation of ordinary and influenced plant leaf can be estimated dependent on variety in temperature, dampness and shading.

The pigments in leaves are in charge of the striking shading changes in the fall. Temperature, daylight and soil dampness all assume a job in how the leaves will look in the fall. Rich daylight and low temperatures after the abscission layer structures cause the chlorophyll to be demolished all the more quickly. We have utilized DHT11 temperature sensor. The DHT11 sensor detects the temperature of the leaf under thought. The parameters that are gathered from the sensor are sent to the cloud stage through the wifi shield associated with the Arduino UNO board. The information which is recorded for investigation in the cloud stage. We at first record the scope of the temperature of a sound leaf. Afterward, if the temperature of the leaf under thought does not fall into that run, at that point the leaf is said to be unhealthy.

Changes in the shade of plant tissue are a typical indication of plant disease. Regularly these shading changes are realized by the yellowing of typical green tissue because of the annihilation of chlorophyll or an inability to frame chlorophyll. Such suppression of leaf shading might be finished or halfway. The shading sensor detects the shade of the leaf under thought which is another parameter that is being utilized to decide if the leaf is either disease or safe.

As indicated by the normal for rural data stream, from the side of innovation, on account of the characteristics of by and large sense, solid exchange and wise procedure, IOT starts to turn into the primary strategy for information securing and transmission and would turn into a significant innovation over a few sorts of sensors to gather, investigate, transmit and deal with the entire information identified with plant illness and creepy crawly pests. The sensor is a significant innovation for information obtaining, and is principally used to catch some portion of the current information, relate and synchronize these information, examine them, lastly, does a receptive movement without client mediation. The parts of a (remote) detecting hub incorporate the accompanying: detecting and in citation unit (single component or exhibit), preparing unit, correspondence unit, control unit and other application-subordinate units Sensors can be basic point components be multi point location clusters, and it has the capacity of huge scale arrangement, low support, scale capacity, flexibility for various situations.

II. Literature Review

This paper shows minimal effort shading sensors for checking plant development in a research facility. A mechanized framework for estimating plant leaf shading is created to check plant wellbeing status [1] by introducing minimal effort shading sensors for checking plant development in a research center. A computerized framework for estimating plant leaf shading is created to check plant wellbeing status [2]. This paper have exhibited novel calculation for division and programmed distinguishing proof of vermin on plants utilizing picture handling. The proposed system includes decreased computational intricacy and goes for irritation recognition in a nursery domain as well as in a ranch situation too. The whitefly, a bio-assailant which represents a danger to a huge number of yields, was picked as the nuisance of enthusiasm for this paper. The calculation was tried for a few whiteflies influencing various leaves and a precision of 96% of whitefly identification was accomplished. We have exhibited irritation control in rural ranches utilizing picture preparing strategies in MATLAB. Pictures are then exposed to pre-handling, change and bunching [3]. This paper has exhibited IoT Implementation for remote checking of agrarian parameters. Remote framework is created to screen natural conditions in horticulture field like temperature, soil pH, soil wet level and stickiness adjacent to leaf sicknesses discovery [4]. We have introduced small scale controller based auto-water system and nuisance recognition utilizing picture handling. A technique for picture examination can widely connected to farming learning for stretch most extreme security of plants which can at last lead to all the more likely yield the executives and generation [5]. Plant diseases and creepy crawly vermin have transformed difficulty by way of which foundation substantial decrease of equally superiority and amount for crop items [6]. In this research we have introduced plan and advancement of vermin observing framework for actualizing exactness agribusiness utilizing IoT. Pakistan the greater part of the rancher develop sugarcane yet did not get yielding because of bugs and hatchlings in sugarcane. In this proposed structure framework utilized arduino for checking the clamor and temperature [7]. This paper has exhibited arduino based vermin control utilizing constant ecological observing sensors. This paper endeavors to build up a robot fit for performing task of apportioning nuisance control specialists, deterrent shirking for self-direction on the field with no client impedance and make a sterile situation for the ideal development of the harvests in a continuous checked shut condition [8]. This paper has displayed a genuine nature sensor and reasonable assessment calculation for plant acknowledgment. The framework created depends on free and programmable real nature sensors for constant acknowledgment and distinguishing proof of individual weed and harvest plants utilizing numerical calculations and choice models [9]. This research has introduced apple leaf sickness distinguishing proof utilizing hereditary calculation and connection based element determination strategy. A shading change construction of information RGB picture remained planned right off the bat and after that RGB structure has been changed over to HSI. YUV as well as dim structure. Foundation has evacuated and afterward the illness spot picture was sectioned with district developing calculation (RGA). At last, the sicknesses were perceived by SVM classifier [10]. The brisk progression of new advances and the changing scene of the online world. Web of Things (IoT), Internet of All, cloud-based plans) give a novel opportunity to making modernized and mechanical systems for urban developing, cultivating, and officer administration. Creative advances in machine vision, overall arranging structures, laser advances, actuators, and mechatronics have engaged the improvement and use of mechanical systems and smart advances for precision agribusiness. In this, we present and review mechanical applications on plant pathology and the board, and rising agricultural developments for intra urban cultivating. Nursery impelled the official's structures and headways have been essentially made in the latest years, joining IoT and WSN (Wireless Sensor Network). AI, machine vision, and AI (Artificial Intelligence) have been utilized and associated in agriculture for robotized and mechanical developing. Understanding advancements, using machine vision/learning, have been made not only for planting, water framework, weeding (to some degree), pruning, and procuring, yet moreover for plant ailment acknowledgment and recognizing confirmation. Regardless, plant disorder recognizable proof still addresses a fascinating test, for both abiotic and biotic weight. Various affirmation procedures and advances for perceiving plant disease reactions have been successfully developed; still, the vast majority of them require a controlled area for data getting to dodge false positives. Man-made intelligence techniques (e.g., significant and move learning) present promising results for improving picture dealing with and plant reaction recognizing evidence. Before long, demonstrative expressness is a test for microorganism control and should drive the improvement of mechatronics and mechanical responses for ailment the board [11].

III. Materials and Methodology

The proposed framework comprises of temperature, dampness, and shading sensors for gathering information from plant leaves dependent on variety in temperature, mugginess and shade of plant leaves. The information gathered from the leaves comprises of current ecological variables like temperature, moistness and shading. The progressions that a plant experiences are caught by the

Temperature moistness and shading sensors and dissected with the Arduino programming. The information gathered with temperature, moistness as well as shading sensor are specified for Arduino UNO unit after that data remains conveyed the ranchers. The framework utilizes WiFi shield so as to send the information from the host framework to the cloud stage for examination. The gathered information in the cloud stage is then contrasted with the dataset all together with identify whether the leaf under thought is typical or influenced. The Figure.1 indicates schematic chart of the proposed work. Step by Step explanation of figure 2.

Fig. 1. Plant Leave Affected by Disease.

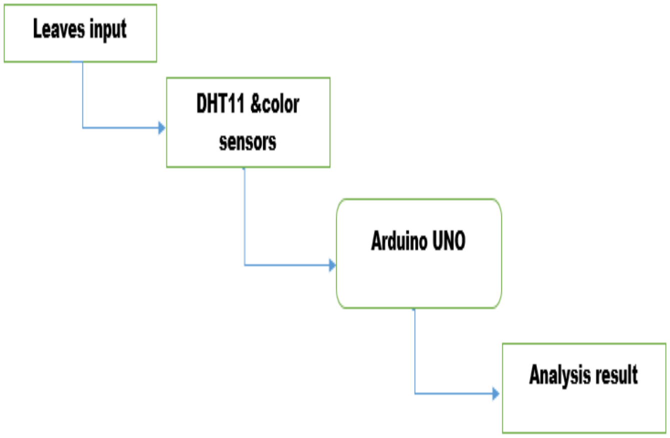


Fig. 2. Proposed Methodology.

1) Information obtaining: Here we accept tests of various leaves as the information. These leaves are then detected by the sensors to decide various parameters dependent on which it is perceived to be sound or infected.

2) Temperature sensors: The DHT11 is a fundamental, ultra simplicity modernized temperature sensor. It uses a capacitive moisture sensor and discharges a propelled sign on the data stick (no straightforward information pins required.

3) Humidity sensor: according figure 3 The DHT11 is a key, ultra simplicity propelled suddenness' sensor. It uses a capacitive moistness sensor and a thermistor to measure the incorporating air, and discharges an automated sign on the data stick.

4) Color Sensor: according figure 4 The TCS3200 is a programmable shading light-to-recurrence converter/sensor. The sensor is a solitary solid CMOS incorporated circuit that consolidates a configurable silicon photodiode and a current-to-recurrence converter.

5) Aurdino: according figure 5 The Arduino United Nation Organization is an comprehensively use open source controller board dependent on the ATmega330P microcontroller as well as prepared by Arduino.cc.

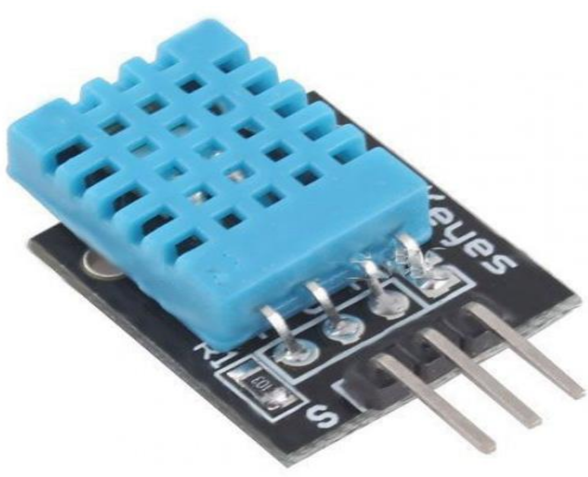


Fig. 3. DHT11 for Humidity and Temperature.

Fig. 4. Tcs3200 for get Color from Leaf.

Fig. 5. Arduino used for Decision.

6) Cloud stage: Here we utilize "ThingSpeak" cloud stage to send the detected information to the cloud. This information sent is plotted against the diagram to see the adjustment in the temperature, mugginess and the shading. Contingent upon the information that is plotted against the diagram we check whether the qualities fall into a similar range. On the off chance that they do as such, at that point the leaf is sound or else it is sick

Algorithm 1: Identification of plant disease using temperature and color sensor

Input: leaf (infected or Normal)
Output: Normal or diseased plant leaf
Description: Given temperature range for the leaf to be healthy is 15-30$_0$C
Start
Step 1: Get leaf for acquisition.
Step 2: measure temperature and color of the leaf using the DHT11 and TCS3200
sensor.
Step 3: calculate the color and temperature
if (minimum range < temperature< maximum range AND minimum range < color < maximum range)
Display "Leaf is Normal"
else

Display “Leaf is Diseased”
Stop

IV. Conclusion

In this paper, a framework is created to decide to the nature of the leaves. The proposed strategy utilizes the sensor gadgets to recognize the parameters like temperature, stickiness and shade of the leaves, which are then contrasted with the informational index with check whether the gathered qualities falls in to the range determined in the informational collection. The proposed model can be utilized in various territories by ranchers, industrialists, botanists, nourishment designers and doctors. The roads for further work here is the point to utilize the picture handling methods alongside the proposed framework to make it progressively proficient and furthermore exact to decide the qualities and to characterize whether the leaves are unhealthy or sound. To assemble an all-encompassing form of the framework, we can utilize the picture preparing system that identify the sort of the sickness the leaf is influenced with and groups the various infections among the leaves. Here we can fabricate a mechanized framework with the goal that it is valuable for the enormous scale creation and furthermore helps in early discovery of the sicknesses that helps the customers for the better execution and upgrades the harvest yield.

The proposed framework is constrained to just identify whether the leaf under thought is solid or infected. This can be additionally done for perceiving the sort of infections in the leaves and arrangement of those ailments. We have constrained our work to just to the temperature, mugginess and shading parameters of the leaves. This can be additionally upgraded by applying different sensors and consolidating with picture preparing ideas. The other confinement is that the decided qualities for the considered parameters are not exact. We have taken the scope of qualities for those parameters and the range may fluctuate dependent on the climatic conditions.

References

[1] Seelye, M., et al. Low cost colour sensors for monitoring plant growth in a laboratory. in 2011 IEEE International Instrumentation and Measurement Technology Conference. 2011. IEEE.

[2] Huddar, S.R., et al. Novel algorithm for segmentation and automatic identification of pests on plants using image processing. in 2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12). 2012. IEEE.

[3] Krishnan, M. and G. Jabert, Pest Control in Agricultural Plantations Using Image Processing. IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), 2013. 6(4): p. 68-74.

[4] SV, M.G. and S. Galande, Measurement of NPK, temperature, moisture, humidity using WSN. 2015.

[5] Rahman, S.A., et al., Reaction Decoder Tool (RDT): extracting features from chemical reactions. Bioinformatics, 2016. 32(13): p. 2065-2066.

[6] Hiary, H., et al., Fast and accurate detection and classification of plant diseases. International Journal of Computer Applications, 2011. 17(1): p. 31-38.

[7] Veerapandi, S., S. Gavaskar, and A. Sumithra, A Hybrid Fault Tolerance System for Distributed Environment using Check Point Mechanism and Replication. International Journal of Computer Applications, 2017. 975: p. 8887.

[8] Sai, V.V. and T. Hemalatha. Computational methods for simulating soil parameters using electrical resistivity technique. in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2017. IEEE.

[9] Schneider, F., et al., The effect of deep tillage on crop yield–What do we really know? Soil and tillage research, 2017. 174: p. 193-204.

[10] Chuanlei, Z., et al., Apple leaf disease identification using genetic algorithm and correlation based feature selection method. International Journal of Agricultural and Biological Engineering, 2017. 10(2): p. 74-83.

[11] Ampatzidis, Y., L. De Bellis, and A. Luvisi, iPathology: robotic applications and management of plants and plant diseases. Sustainability, 2017. 9(6): p. 1010.

Performance Analysis of Machine Learning Classifiers for Detecting PE Malware

ABM.Adnan Azmee[1], Pranto Protim Choudhury[2], Md. Aosaful Alam[3], Orko Dutta[4], Muhammad Iqbal Hossain[5]
Computer Science and Engineering, BRAC University
Dhaka, Bangladesh

***Abstract*—In this modern era of technology, securing and protecting one's data has been a major concern and needs to be focused on. Malware is a program that is designed to cause harm and malware analysis is one of the paramount focused points under the sight of cyber forensic professionals and network administrations. The degree of the harm brought about by malignant programming varies to a great extent. If this happens at home to a random person then that may lead to some loss of irrelevant or unimportant information but for a corporate network, it can lead to loss of valuable business data. The existing research does focus on some few machine learning algorithms to detect malware and very few of them worked with Portable Executables (PE) files. In this paper, we mainly focused on top classification algorithms and compare their accuracy to find out which one is giving the best result according to the dataset and also compare among these algorithms. Top machine learning classification algorithms were used alongside neural networks such as Artificial Neural Network, XGBoost, Support Vector Machine, Extra Tree Classifier, etc. The experimental result shows that XGBoost achieved the highest accuracy of 98.62 percent when compared with other approaches. Thus, to provide a better solution for this kind of anomalies, we have been interested in researching malware detection and want to contribute to building strong and protective cybersecurity.**

***Keywords*—Malware detection; machine learning; data protection; XGBoost; support vector machine; extra tree classifiers; artificial neural network**

I. Introduction

Malware or malicious software, a dangerous computer code, intended to disturb, cripple or take control of the computer system without the approval of the user. It takes advantage of the technical faults or vulnerabilities in the operating system, hardware, and software. Malware is frequently used to take assets from the PC or attempting to take some significant data, records or cash from an individual. Malware has been there on the internet for quite a long amount of time and "Brain" is considered to be the first virus in the history on the personal computer (pc). It was originated by certain Pakistani young boys in their late teens for their motivations and purposes. Amjad Farooq Alvi was the mind behind the coding section of Brain [1]. In this 21st century, both malware and its spreading rate are growing rapidly along with the advancement of technology. Malware and viruses are also used in cyber warfare between countries. Iran and Saudi Arabia have been in cyber warfare for more than 10 years. However, every malware is not disastrous but it can cause certain limits of distress like it can cause the laptop or computer to slow down or run at a slow pace and can also cause irritating conduct like creating a series of pop-up advertisements.

Antivirus is often unaware of any new virus or malware that is being spread through the internet and by the time a solution comes, users have already been affected by the new virus. In addition, this can lead to the loss of useful information and money. Saving personal data is everyone's concern, but in this era of technology, it seems quite impossible. This malicious malware is the reason for losing billions of dollars as well as data. Cybercriminals will take an expected thirty-three billion records by 2023 as indicated by a recent report from Juniper Exploration [2]. Every year billions of data are being stolen from various sites and some of them are used even for bad purposes. Almost sixty million Americans have been influenced by data fraud as per a 2018 online overview by The Harris Survey. A similar overview demonstrates about fifteen million purchasers experienced data fraud in 2017 [3]. These attacks are also causing great financial loss. In the Accenture report, they say that a company on average costs us dollars due to malware attacks [4].

Much research has been conducted on making new techniques and strategies to assemble, study and ease noxious code as malware is spreading at an alarming rate on the web. Unfortunately, current host-based detection methods undergo weak detection models. These models focus on the highlights of a particular malware occasion and are regularly effectively evadable by obfuscation or polymorphism. Additionally, finders that check for the nearness of a grouping of framework calls displayed by a malware example are frequently evadable by framework call reordering. To address the inadequacies of weak models, a few powerful discovery approaches have been suggested that expect to recognize the conduct shown by a malware family. Albeit promising, these methodologies are tragically too hindered to even think about being utilized as constant finders on the end host, and they regularly require lumbering virtual machine innovation. In our thesis, we carry out a comparative study on different machine learning classification algorithms in detecting malware and benign PE (Portable Executable) files. We used a dataset from Kaggle, which was built using python library (PE files) and contains malicious and benign data of PE files. At first from the given dataset we perform data pre-processing and then implement them on the machine learning algorithms we select to work with. In our research, we used several machine learning algorithms to assemble several classifiers and then we used those classifiers to detect the PE malware and after that, we compared the performance of classifiers in accurately identifying malware. We compared different machine algorithms and tried to figure out which algorithm has the highest accuracy to detect the malware.

II. Literature Review

In [5], the authors worked on portable executables and tried to figure out which of the files are malware and created an integrated feature set based on raw and derived inputs. Moreover, they had used only six algorithms to compare between integrated and raw feature set. Furthermore, they used the 10 fold cross-validation technique. They created two datasets with the help of virus-share, Windows XP and Windows 7. One of them had two thousand seven hundred twenty-two malware and two thousand four hundred eighty-eight benign data and another one had one hundred twenty-nine malware data and thirty benign data. Lastly, they changed the feature numbers several times. In our paper, we use more algorithms than them, and our dataset is more ethical because we collected it from Kaggle. And, we have more data than their dataset.

In [6], Singhal made a model that was intended to help in the network security of enterprises. He did it because they felt that the signature-based antivirus system could be enough for household purposes but might be a threatening issue for networks of enterprises. However, he used only three algorithms to compare with their model and worked on only five thousand data. His model gave ninety-eight percent accuracy but only worked well for enterprise networks, not with the personal computers of home. In addition, he did not mention any processing techniques like PCA, LDA but we used some of those techniques in our work. We are also ahead of him in comparison to the highest accuracy rate of algorithms and the number of data in the dataset.

A PE malware detection system is created by the authors of the paper [7]. Actually, they worked in three steps, these are feature extraction, selection, and classification. For feature extraction, pefile, which is a python module had been used. However, they only used five hundred fifty-two data, which is much less than us. Moreover, they used the chi-square test, which found the relation between features with statistics and eventually works well with the small amount of data. If there are much data, it can lead to erroneous conclusions whereas our model did well with a large amount of data. We use a variety of algorithms like decision trees, boosting, and neural networking. Lastly, their accuracy rate was 97.25%, we found more accuracy in our work even with a large amount of data.

Another real-time malware detection work was done on [8] by authors. They extracted only 35 features, again less than our number of features. They used only four algorithms; CNN, MLP, SVM and random forest and netmate technique for feature extraction. It is another paper using fewer algorithms and got less accuracy rate than us. Each of their algorithms gave an accuracy rate of more than eighty-five percent. Whereas in our work, we got a more than eighty-five percent accuracy rate in several algorithms. And they did not use any feature reduction technique.

In [9], Merabet *et al.* research had been done on machine learning malware detection. In the early stage of our work, we got different ideas from this paper. According to them, the heuristic machine learning is better than the signature method which cannot stop new malware and dynamic method which has time and space complexity They made the comparison on several steps which are necessary for machine learning malware detection. Firstly, they compare among different feature extraction techniques like signature-based, dll function call, binary sequence, assembly sequence, pe file header, machine activity matrix, entropy signal. Secondly, they showed their survey on feature selection methods like information gain, redundant feature removal, principal component analysis, random forest, self-organizing feature map, wavelet transform. Lastly, they discussed three algorithms, these are supported machine vector, random forest, and artificial neural network and they compare accuracy rate based on feature selection and algorithms.

A hybrid machine learning technique is used for malware detection in [10] by the authors. They had the target of tracing and categorizing malware and had a better accuracy rate. For gaining that, they collected 5.05 GB benign along with 1.28 GB malware data and extract of n-gram and pe type of data. They only used three supervised classifiers (J48, Random Forest, Naive Bayes) and one unsupervised classifier (Self-organizing feature map). They used the information gain technique for feature selection.

And, in [11], Anderson *et al.* created a big dataset and opened it for all for research purposes. Their dataset contained three hundred thousand malicious data, three hundred thousand benign data, and three hundred thousand data which were not labeled. For test purposes, they reduced it to one hundred thousand malicious data and three hundred thousand benign data and make another dataset. However, they used eight types of features, only used LightGBM and proved that they gave a better performance than another deep learning model, malcov.

III. Proposed Approach

The workflow of our approach is given below Fig 1 to distinguish obscure malware by using machine learning. Our approach includes several steps to make it more accurate, efficient and effective. Our approach includes preprocessing of the dataset, feature selection, Training/Testing data in machine learning classification algorithm to detect malware or benign. After applying the classification algorithm we analyzed and visualized the result.

The dataset makes the machine learning training feasible. The dataset is used to train the model for performing various actions, to work automatically. The training dataset is a dataset in which the machine learning algorithm has trained and the dataset we use to validate the accuracy of our model is called testing dataset. Datasets can be built, downloaded from the website. Our dataset was collected from Microsoft Kaggle which contains nineteen thousand six hundred eleven samples. The specialty of our dataset is too many features. Our dataset contains seventy-nine features.

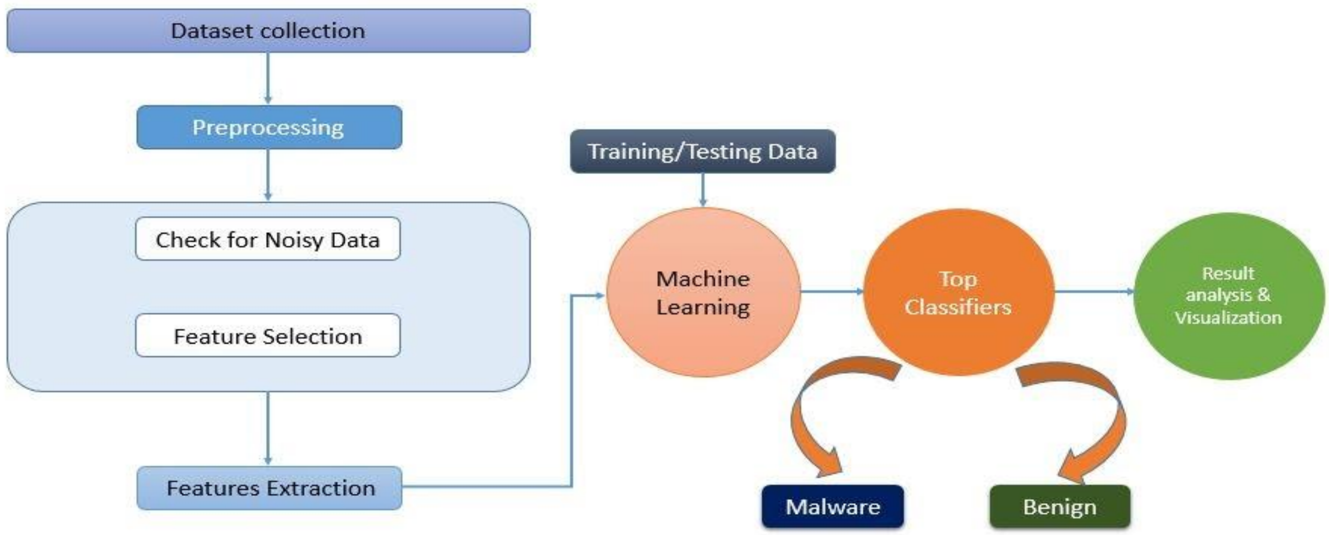


Fig. 1. The Workflow of our Approach.

Data preprocessing is an information mining strategy that is used to change the unrefined information in an accommodating and appreciable format. One of the data preprocessing steps is PCA (Principal Component Analysis) which we used in our approach. PCA is a dimension reduction technique to avoid overfitting. It takes a dataset and "rotates" it, taking the original axes characterized by the original factors, and making new axes that are linear combinations of the old data. The linear combination exact is picked with the end goal that each progressive segment amplifies fluctuation along that new dimensions. It is a method of summarizing the data. Another preprocessing step we used in the dataset is that we used standard scalar. Standardization of a dataset is a typical necessity for some, machine learning estimators. They may carry on severely if the individual features do not basically look like standard usually scattered information, For example, numerous elements utilized in the target capacity of a learning algorithm accept that all features are based on zero and fluctuation in a similar request It will transfer data that its distribution will have a standard deviation of one and mean value zero.

Noisy data is pointless data. Any information that has been received, stored, or changed in such a way, that it can't be perused or utilized by the program that initially made it tends to be depicted as noisy. We checked for any noisy data in our dataset and we found nothing.

Feature Selection is where you consequently or physically select those Features which contribute most to your prediction variable or yield in which you are keen on.

Having unimportant Features in data can minimize the accuracy of the models and cause the model to learn reliant on immaterial Features. Feature selection can reduce overfitting, training time [12]. In our dataset, we have seventy-nine features. Among them, we worked with seventy-seven features. We dropped the target column and another column “Name”. We used Weka for feature selection as shown in Fig. 2.

Weka which stands for Waikato Environment for Knowledge Analysis is a collection of information analysis and visualization tools, exhibited by the University of Wakito. In weka in the attribute evaluator, we used ClassifierAttributeEval and for the searching method, we used Ranker. Ranker ranks each attribute and returns a score. For attribute selection mode we used 10 fold cross-validation. However, we used all the features of our work.

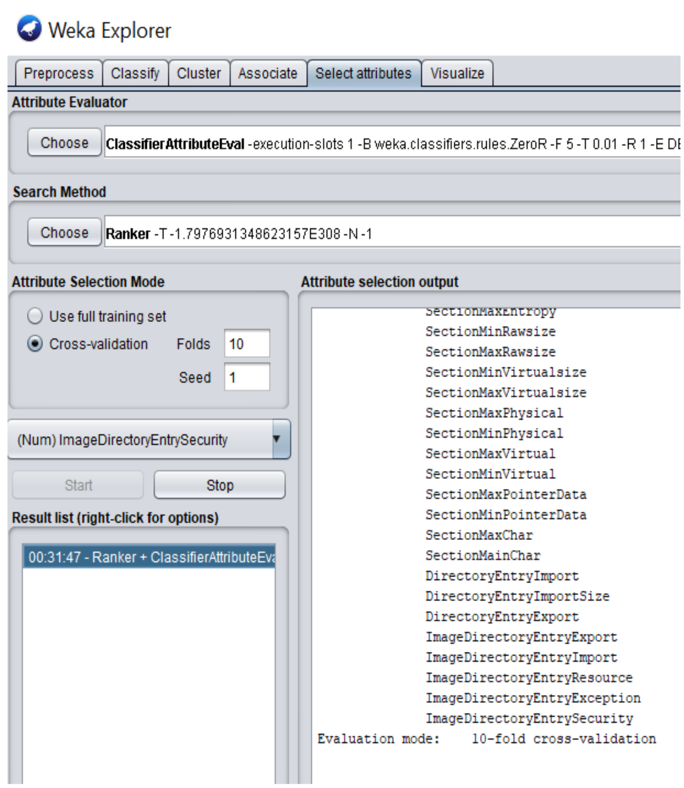


Fig. 2. Feature Selection by Weka.

After the feature selection process, the next step is training and testing of machine learning classifiers. We split our dataset into training and testing sets. Then we train and test the dataset in some classification algorithms. Training data are used to fit and tune the models. Test data are represented as unseen data to evaluate the models [13]. Test data is utilized to perceive how well the machine can foresee new answers dependent on training. In our dataset, we used eighty percent data as training data and twenty percent as testing data. After splitting the dataset we train our model. We have used nine classification algorithms for our approach. The nine classifiers are

1) Logistic Regression
2) K-Nearest Neighbor (KNN)
3) Random Forest
4) Adaboost
5) Support Vector Machine (SVM)
6) Decision Tree
7) XGBoost
8) Artificial Neural Network (ANN)
9) Extra Tree Classifier

These classification algorithms were used with particular laws to detect whether it is malware or benign.

After receiving the outcome, they were analyzed and visualized. As we used nine classification algorithm we analyzed their accuracy, Precision, Recall, True Positive Rate (TPR), False Positive Rate (FPR), f1-Score, Support, Confusion Matrix. For the visualization part, we have used Heatmap, Distribution Graph, Correlation Matrix, Pie plot, Counterplot.

IV. Dataset Description

Each executable document has a typical arrangement called Common Object File Format (COFF), an organization for executable, object code, shared library PC records utilized on Unix frameworks [14]. Also, PE (Portable Executables) design is one such COFF format accessible today for executable, object code, DLLs, FON font documents, and core dumps in thirty-two bit and sixty-four-bit versions of Windows operating systems. PE header actually holds the required data for the operating system to run the installation files like exe type data [15]. This comprises dynamic library references for linking, API export, and import tables, resource management data and TLS data. The Data Structures that are on the disk are the exact data structures that are being used in the memory. Instead of direct mapping of PE files into memory as a single memory-mapped file, the Win32 loader decides which portions of the PE files need to be mapped. Our dataset is from Kaggle and they used the python pefile module to generate information of pe header and sections of pe file. PE format is of two types. PE format is a well-known windows 32-bit format and PE+ is a well-known windows 64-bit format. The PE File structure Section symbolizes a portion of memory that contains either code or data and Section table contains numbers of sections. PE header holds the information that the operating system needs to know to run the executable, for example: (.exe) type files. Lastly, the DOS Mz header contains the offset of the PE header and Dos Stub prints whether the executables will run or not.

Here, Fig 3 represents the structure of a PE file which is given.

We have used seventy-seven columns as input because the column “name” for input has string values and the column “malware” is our target or the output column. In TABLE I, only a few important descriptions of columns are given.

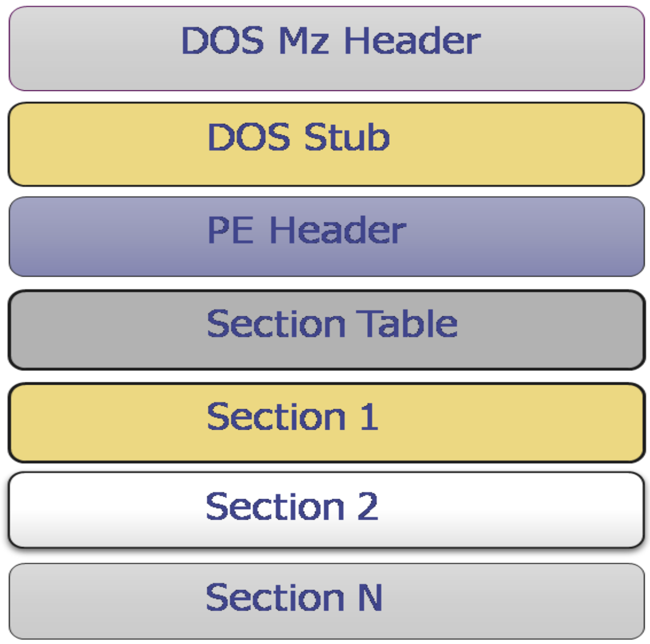


Fig. 3. Structure of PE Header.

TABLE. I. Dataset Description

	COLUMN NAME	DESCRIPTION
1	Magic	We need to know whether the executable image is of thirty-two bits or sixty-four bits and the magic field tells this accurately.
2	AddressOfEntryPoint	This holds the RVA (Relative Virtual Address) of the Entry Point (EP) of the module and is normally found in the text section.
3	BaseOfCode	BaseOfCode holds the RVAs of the beginning of the code.
4	BaseOfData	BaseOfData holds the data section of the beginning of the code.
5	ImageBase	Executable file needs to be memory-mapped to an exact location in memory and ImageBase is the address where it is done. 0x10000 is the default ImageBase for an executable in Windows NT and 0x400000 is the default ImageBase for DLL.
6	SectionAlignment	SectionAlignment designates to the alignment of the sections of PE in the memory.
7	FileAlignment	FileAlignment indicates the alignment of the sections of PE in the file.

V. Algorithms

We are using supervised machine learning which refers to the events when we have all information are labeled or in another way, we can say we know how all information is classified. Classification and regression are types of supervised learning where we aware of the classes of classification and values in regression.

A logistic regression algorithm is used to predict discrete or categorical values. It is mainly a classification algorithm that is used in cases like fraud detection, email spam detection, etc. where we have to make decisions between yes and no. It predicts the likelihood of the event of an occasion by fitting information into the logit function. Generally, a logistic regression model calculates the class membership probability for one of the two categories in the data set [16]. However, the Logistic curve is not a straight curve like linear regression. It is known as the sigmoid curve and here probability.

$$p = 1/\, 1 - e\hat{}\, - z \tag{1}$$

where z= mx+c. This equation of probability ensures that the predictor will be between 0 and 1. In our work after preprocessing by standard scaler, we used the logistic regression classifier.

K-Nearest Neighbor (KNN) is the simplest machine learning algorithm that is used for both classification and regression models. Whenever the model is fed with testing data it finds the distance of that point with every other point in the training data. Then it finds the nearest k members for that point. The implementation of KNN can be done by following some steps which are given below:

- Load the information.
- Initialize the estimation of k.
- For getting the anticipated class, emphasize from 1 to add up to the number of training information points.
- Calculate the difference between test information and each row of training information. Here we will utilize Euclidean separation as our separation metric since it's the most prominent technique. For example, if we have two points P1(X1, Y1) and P2(X2, Y2) then

$$d(P1, P2) = \sqrt{(Y2 - Y1)^2 + (X2 - X1)^2} \tag{2}$$

Random Forest, like its name induces, is various decision trees joined into a single model. How Random forest capacities are by at first building trees and afterward accumulating them. The more the data the better the outcome as we merge all the more learning. As all the decision trees are merged, the figures will be closer to the mark. When looking at is done with substitutions, it draws the training set for current trees. After it's fulfillment, around thirty-three percent of the cases are chosen not to take any other potentially detrimental action for the model. As trees are added to the forest, the running fair check of the classification can be dictated by using the out-of-sack data. Central purposes of using random forest include dealing with both course of action and backslide issues, working with a categorical and fast and tenacious variable, normally handles missing regards, not requiring scaling and is less influenced by noise.

Boosting, in general, is a sequential process where we create multiple weak classifiers into a big ensemble strong classifier. It helps to improve the learners by focusing on areas where the system is not performing well. One of the best algorithms in this sector is AdaBoost, also known as Adaptive Boosting. The algorithm mainly increases the weight of certain training data inputs by using ensemble learning. There are some steps on how AdaBoost performs:

- First of all, it initializes weights to all training points. At initialization, every point that is used to train the classifier will be assigned a weight which is equal to the reciprocal of the total number of points that are present in my training data set.
- Then it calculates the error rate for each weak classifier that is present by growing multiple decision stumps and pick the decision stumps with the lowest error rate.
- The voting power of the weak classifier is computed and then appended the classifier into the ensemble classifier.
- The next step is to update the weight of every training point based on whether it was classified correctly or incorrectly by the previous classifier. The weight goes up for a point if it was classified incorrectly and goes down if it was classified correctly.

Support Vector Machine is a supervised machine learning algorithm that can be used for both classification and regression. The goal of SVM is to find the hyperplane which divides the two classes of the data. Under the assumption that two classes are considered the training set D (input space) of N pairs (xi, yi), i = 1,..., N, which is used during the training process can be defined as follows [17]:

$$D = \{ (xi, yi) | xi \in R^n, yi \in \{-1, +1\} \} \tag{3}$$

Given a training dataset the SVM algorithm searches for a plane (a hyperplane for n > 3) in the input space that separates the positive samples from the negative ones. In the original SVM model, all hyperplane in Rn is parameterized by a vector named w and constant b.

$$w^T x + b = 0 \tag{4}$$

For hyperplane (w, b) defined in (2) that separates the data, we can formulate the classification rule.

$$h(x) = sign\ (w^T x + b) \tag{5}$$

Here h(x) will correctly classify the sample data.

The decision tree algorithm can be utilized to take care of both classification and regression issues. We can speak to any boolean capacity on discrete characteristics utilizing the Decision tree. There are a few presumptions we have to settle on while utilizing decision trees. At first, we need to consider the root hub as the training set. Feature values are liked to be straight out. On the off chance that the qualities are continuous, at that point, they are discretized preceding the structure of the model. Based on quality qualities records are conveyed recursively. We utilize factual strategies for requesting qualities as root or the inside hub. In Decision Tree, the significant test is to recognizable proof of the characteristic for the root hub at each level. This procedure is known as attribute selection. We have two famous property choice measures:

- Information Gain.
- Gini Index.

XGBoost is an algorithm that has as of late been commanding applied machine learning and Kaggle competitions for organized or unthinkable information. XGBoost is a usage of gradient boosting decision trees intended for speed and execution.XGBoost modifies the boosting algorithm based on GBDT(Gradient Boosting Decision Trees). This algorithm is made out of numerous regression trees, and the last outcome is an added substance mix of the decision results of all subtrees. In any case, the glaring exclusion of GBDT is the need to use the leftover mistake of the n-1th tree when training the nth tree, which makes GBDT hard to be conveyed on the disseminated framework. regular item is added to the loss function to avoid overfitting which is given below:

$$\Omega(f_t) = \square T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j{}^2 \quad (6)$$

Artificial Neural Network Algorithm (ANN) is developed from the idea of how the neurons of human brains are connected, and how the nervous system performs its work. Artificial Neural Network has many uses and classification is one of them. It can perform well for large datasets; instead of taking the entire dataset it takes data samples. Basically, the artificial neural network has three different layers, the first one is the input layer, where data are given as input, the second one is the hidden layer and the final one is the output layer. Depending on the number of hidden layers varies. ANN displaying begins with randomly appointed weight coefficients. At that point, a lot of information designs are nourished forward over and again, and loads of the neurons are changed until the yield coordinates intimately with the real values. In our model, we used two hidden layers. We used Rectified Linear Unit (ReLU) and Sigmoid function as the activation function in our model. After the model is defined we compiled it, we used cross-entropy as the loss function. For optimization, we used Adam optimization algorithm that is the extension of a greatly used optimization algorithm stochastic gradient descent.

The Extra-Tree technique (representing extremely randomized trees) was proposed with the primary goal of further randomizing trees working with regards to numerical information features, where the decision of the ideal cut-point is answerable for an enormous extent of the fluctuation of the incited tree. The Extra-Trees algorithm constructs a gathering of unpruned choice or relapse trees as per the traditional top-down strategy. Its two principal contrasts with other tree-based outfit techniques are that it parts nodes by picking cut-focuses completely indiscriminately and that it utilizes the entire learning test (as opposed to a bootstrap copy) to develop the trees. The Extra-Tree has two parameters: K, the number of properties arbitrarily chose at every node and nmin, the least example size for splitting a node. It is utilized a few times with the (full) unique learning test to create a troupe model (we mean by M the number of trees of this group). The forecasts of the trees are collected to yield the last expectation, by the greater part vote in characterization issues and number juggling normal in relapse issues. From the predisposition change perspective, the justification behind the Extra-Trees strategy is that the unequivocal randomization of the cut-point and property joined with gathering averaging ought to have the option to diminish change more emphatically than the more fragile randomization plans utilized by different strategies. The use of the full unique learning test as opposed to bootstrap copies is persuaded to limit predisposition.

VI. Result Analysis and Visualization

A. Data Visualization

In order to work with any dataset, it is very important to visualize it, python has a bunch of libraries with the help of which we can easily visualize the data. We used the python visualization library to plot different plots. Fig 4 represents the Heatmap and Fig 5 shows the Countplot of the dataset.

We used the python seaborn library to plot the heatmap. Heatmap is actually a colored tabular matrix where the color of each cell depends on the data contained in it. It shows the correlation. We also used count plot. A count plot is used to visualize the number of items present in each category in the form of a rectangular bar. We used the seaborn visualization libraries count plot to visualize our data. In our data 0 stands for benign and 1 stand for malware.

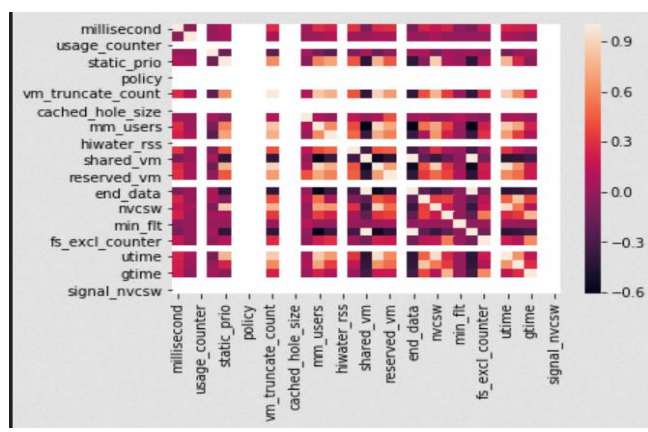


Fig. 4. Heatmap of the Input Data.

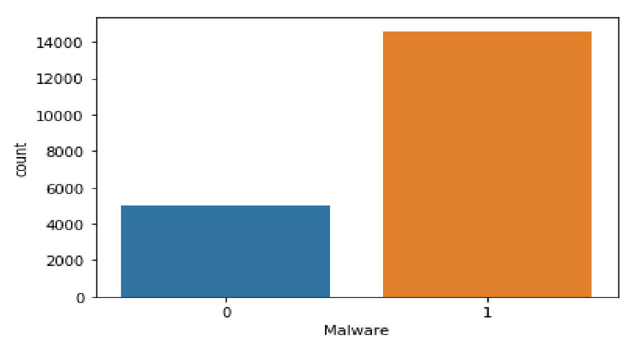


Fig. 5. Count Plot of our Dataset.

B. Result Analysis

In this part, we are going to analyze the result with the help of different evaluation metrics such as classification report, confusion matrix, roc curve, accuracy score, etc.

We used nine classifiers to classify the malware and benign data, they are Logistic Regression, K-Nearest Neighbor (k-NN), Random Forest, AdaBoost, Support Vector Machines (SVM), Decision Trees, XGBoost, Artificial Neural Network (ANN) and Extra Tree Classifier. After preprocessing we split the data into training and test set. We trained the classifier with the training set and then we tested them using the test data. We evaluated the performance of the classifiers with the help of different metrics. Different metrics used by us for evaluation are discussed as follows:

C. Confusion Matrix

One of the most widely used methods of evaluating the machine learning algorithm is the confusion matrix. TABLE II shows the different combinations of Confusion Matrix.

Here, True Negative (T.N) means that the algorithm predicted negative and the result is actually negative, False Positive (F.P) means that the algorithm predicted positive but the result is actually negative, False Negative (F.N) means that the algorithm predicted negative but the result is actually positive and True Positive (T.P) means that the algorithm predicted positive and the result is actually positive.

True Positive Rate (TPR): TPR is also known as recall. It shows the ability of our classifier to detect malware

TABLE. II. CONFUSION MATRIX

Confusion Matrix	Negative (Predicted)	Positive (Predicted)
Negative (actual)	True Negative (TN)	False Positive (FP)
Positive (actual)	False Negative (FN)	True Positive (TP)

TPR = True Positive / (True Positive + False Negative) (7)

False Positive Rate (FPR): FPR shows that the possibility of benign files wrongly classified as malware.

FPR = False Positive/ (False Positive + False Negative) (8)

D. Classification Report

The classification report is another evaluation metrics for the evaluation machine learning algorithm. We used the python sklearn libraries classification report to show the classification report of different algorithms. With the help of the classification report, we calculate the parameters like Precision, Recall, F1 score, Support, etc.

Precision is the ratio of correctly predicted positive to the total predicted positive sample. The recall is the ability of an algorithm to find all positive samples. F1 Score is the weighted average of Precision and Recall. support is the number of samples of the true sample that lie in that class i.e. number of occurrences of each class.

E. ROC Curve

Receiver Operating Characteristic curve (ROC Curve) is used to plot the true positive rate against the false-positive rate. With the help of ROC Curve, we can find the capability of a classifier to differentiate between different classes. The Area Under Curve (AUC) is the area under the ROC Curve, with the help of AUC score we get the idea of how well the model is performing.

F. Accuracy Score

An accuracy score is the most common metric for evaluating a model. We used sklearn Accuracy score metrics to find the accuracy score for different algorithms.

Accuracy score= (TP+TN)/ (TP+TN+FP+FN) (9)

TABLE III represents the result of the experiment, which is given.

TABLE. III. COMPARISON BETWEEN DIFFERENT ALGORITHMS EVALUATION METRICS

Applied Algorithms	Accuracy		Precision	Recall	f1-score	support	TPR	FPR
Logistic Regression	96.3%	0	0.92	0.93	0.93	990	0.974	0.075
		1	0.98	0.97	0.98	2933		
KNN	95.9%	0	0.92	0.92	0.92	986	0.974	0.083
		1	0.97	0.97	0.97	2937		
Random Forest	98.1%	0	0.94	0.98	0.96	953	0.994	0.057
		1	0.99	0.98	0.99	2970		
AdaBoost	97.2%	0	0.94	0.95	0.94	976	0.984	0.064
		1	0.98	0.98	0.98	2947		
SVM	96.3%	0	0.91	0.94	0.93	968	0.979	0.077
		1	0.98	0.97	0.98	2955		
Decision Tree	97.2%	0	0.95	0.95	0.95	996	0.980	0.051
		1	0.98	0.98	0.98	2927		
XGBoost	98.6%	0	0.96	0.98	0.97	975	0.994	0.037
		1	0.99	0.99	0.99	2948		
Artificial Neural Network	98.0%	0	0.98	0.94	0.96	995	0.993	0.057
		1	0.98	0.99	0.99	2928		
Extra Tree Classifier	95.9%	0	0.93	0.91	0.92	1014	0.969	0.069
		1	0.97	0.98	0.97	2909		

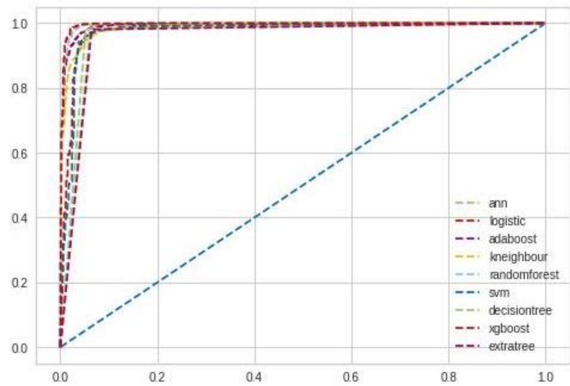


Fig. 6. Combined ROC Curve of All Algorithms.

At first, we used classification algorithms like logistic regression, K-nearest neighbor and SVM (Support Vector Machine) to classify our data. For small to medium size, tabular data most of the time decision tree-based algorithm performs best. Since we are using tabular data, we used many decision tree-based algorithms like decision tree, then random forest which basically is an ensemble method that consists of multiple decision trees. We used an extra tree classifier that works like a random forest but is much faster. Boosting algorithm like AdaBoost and XGBoost is used. We also used Artificial Neural Network to classify our data. Among the algorithms, XGBoost performs best due to its features like regularization, cross-validation, tree pruning, etc. Fig 6 represents the combined ROC curve of all algorithms.

Here the combined ROC curve shows the ratio of true positive rate and false positive rate of different algorithms. Here the AUC (Area Under Curve) for XGBoost is the highest that is 0.99.

VII. Conclusion and Future Works

In our work, we used multiple methods such as KNN, logistic regression, SVM, XGBoost, Decision Tree, etc. classifiers are used along with the artificial neural network. The dataset used in the proposed model had 19611 samples and the dataset was labeled. All the nine methods are compared that were used to detect the malware. After comparing all the three methods, we found that XGBoost got the highest accuracy of 98.6% whereas TPR is 0.99 and FPR is 0.037 with an AUC of 0.99. The second-best accuracy we got from Random Forest was 98.1%.KNN and Extra Tree Classifier got the lowest accuracy which was 95.9%. In any case, completely assess the reasonableness of our methodology, a lot more examination need to direct. While our underlying outcomes are promising, more works are needed to improve the technique and accuracy of our proposed model. For the future, we will consider a more advanced neural network alongside ANN. We are planning to build a hybrid model and an antivirus in the future, in which we are going to implement the hybrid machine learning algorithm. The antivirus would detect the PE file and predict whether they are malware or benign. We hope it would be able to tackle zero-day attacks, which the current antivirus is not able to.

References

[1] J. Harán and J. Harán, "Malware of the 1980s: Looking back at the Brain Virus and the Morris Worm | WeLiveSecurity", WeLiveSecurity, 2019. [Online]. Available: https://www.welivesecurity.com/2018/11/05/malware-1980s-brain-virus-morris-worm. [Accessed: 30- Dec-2019].

[2] “10 cyber security facts and statistics for 2018", Us.norton.com, 2019. [Online]. Available: https://us.norton.com/internetsecurity-emerging-threats-10-facts-about-todays-cybersecurity-landscape-that-you-should-know.html. [Accessed: 30- Dec- 2019].

[3] "LifeLock Official Site | Identity Theft Protection", Lifelock.com, 2019. [Online]. Available: https://www.lifelock.com/learn-identity-theft-resources-how-common-is-identity-theft.html. [Accessed: 30- Dec-2019].

[4] "Cybersecurity Consulting Services | Accenture", Accenture.com, 2019. [Online]. Available: https://www.accenture.com [Accessed: 30- Dec-2019].

[5] A. Kumar, K. Kuppusamy and G. Aghila, "A learning model to detect maliciousness of portable executable using integrated feature set", Journal of King Saud University - Computer and Information Sciences, vol. 31, no. 2, pp. 252-265, 2019. Available: 10.1016/j.jksuci.2017.01.003.

[6] P. Singhal, "Malware Detection Module using Machine Learning Algorithms to Assist in Centralized Security in Enterprise Networks", International Journal of Network Security & Its Applications, vol. 4, no. 1, pp. 61-67, 2012. Available: 10.5121/ijnsa.2012.4106.

[7] "A Chi-Square-Based Decision for Real-Time Malware Detection Using PE-File Features", Journal of Information Processing Systems, 2016. Available: 10.3745/jips.03.0058.

[8] "CSDL | IEEE Computer Society", Computer.org, 2019. [Online]. Available: https://www.computer.org/csdl/proceedings-article/icoin/2018. [Accessed: 30- Dec- 2019].

[9] H. Merabet and A. Hajraoui, "A Survey of Malware Detection Techniques based on Machine Learning", International Journal of Advanced Computer Science and Applications, vol. 10, no. 1, 2019. Available: 10.14569/ijacsa.2019.0100148.

[10] R. Yang, V. Kang, S. Albouq and M. Zohdy, "Application of Hybrid Machine Learning to Detect and Remove Malware", Transactions on Machine Learning and Artificial Intelligence, vol. 3, no. 4, 2015. Available: 10.14738/tmlai.34.1436.

[11] Anderson, H. S., Roth, and Phil, “EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models,” arXiv.org, 16-Apr-2018. [Online]. Available: https://arxiv.org. [Accessed: 10-Dec-2019].

[12] "Feature Selection Techniques in Machine Learning with Python", Medium, 2019. [Online]. Available: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e. [Accessed: 30- Dec- 2019].

[13] "Chapter 6: Model Training with Machine Learning - Data Science Primer", EliteDataScience, 2019. [Online]. Available: https://elitedatascience.com/model-training. [Accessed: 30- Dec- 2019].

[14] "PE Format - Win32 apps", Docs.microsoft.com, 2019. [Online]. Available: https://docs.microsoft.com/en-us/windows/win32/debug/pe-format. [Accessed: 30- Dec- 2019].

[15] "Exploit Development 02—PE File Format 1", Medium, 2019. [Online]. Available: https://medium.com/@MKahsari/exploit-developm ent-02-pe-file-format-1-998c252b5670. [Accessed: 25- Dec- 2019].

[16] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review", Journal of Biomedical Informatics, vol. 35, no. 5-6, pp. 352-359, 2002. Available: 10.1016/s1532-0464(03)00034-0.

[17] Cs.cornell.edu, 2019. [Online]. Available: http://www.cs.cornell.edu/courses/cs578/2003fa/slides_sigir03_tutorial-modified.v3.pdf. [Accessed: 30- Dec- 2019].

Determinants of Interface Criteria Learning Technology for Disabled Learner using Analytical Hierarchy Process

Syazwani Ramli[1], Hazura Mohamed[2], Zurina Muda[3]
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia (UKM)
43600 Bangi, Selangor

***Abstract*—The advancement of technology nowadays is rapidly increasing due to the leveraging availability of learning technology. Due to the rapid change in the availability of technology, it is crucial for disabled learners to select a good technology design that may help them to achieve better academic achievements. Selecting a good design of technology involves a decision making process to choose several designs of learning technology. In general, the abilities, capacities and achievements of disabled learners are lower compared to a normal child. Using a good approach and assisted with the right selection of learning technology may help disabled learners to get a better understanding and achievement in academic matters. In this study, the analytical hierarchy process (AHP) approach was used to determine the best appropriate design of learning technology for disabled learners. Three hierarchy levels made up of criteria, sub-criteria and alternatives were considered. This study finds the best selection design of elements that can be used in the development of learning technology in a classroom of disabled learners.**

Keywords—Selection design; learning technology; disabled learner; decision making; Analytical Hierarchy Process (AHP)

I. Introduction

In this sophisticated era of technology, the availability of technology may help human in many aspects in life especially in education field. In general, the suitability of technology is the main concern for adaptability of human aspects especially for disabled learner. As we known, disabled learner has a limitations and capacity compared to as a normal learner. Disabled learner needs a certain approach, design and methods to adapt on their capability especially in the usage of a technology.

The technology not only focusing on certain area content of teaching aspect but also need to focus on the limitations and difficulties that disabled learner has [1]. Lack accessing appropriate of technology and unsuitable design of website and online learning may become disadvantage to disabled learner [2]. It is shown that suitability aspect of design technology learning need to be considered as a crucial part for development of learning technology for disabled learner.

In the 21st century technology nowadays, technology may provide in assisting disabled learner to help them understand on their learning process. Evolutions development of information technology sector is one of component in 21st century. In the 21st century, modes of electronic are emphasize specially for the disabilities student and risk learners [3]. Action and policies should be emphasized to disabled learner in order to make sure they are not being left behind with others in gaining equality in education. As we known that disabled learner has limitations in terms of cognitive level, lack of the physical movement and psychological activity.

Making a decision with the support of family members, ability to speak in English, writing, reading and finding a job is an essential business of an education for disabled learner [4]. Literally, disabled learner should have a good support system to make sure them in a good track especially in process of learning. They need to assist with a right approach and methods in order to attract their attention and having fun in a learning class environment. In designing of an application with accessible functions easy for disabled learner to interact, understand, accomplish some tasks activities and response to the system [5].

Best pedagogical design in assistive technology is one of the crucial issues that might help student with disability and sensory limitation [6]. A best design with their suitability and approaches may help to overcome their limitations and level of capability that they have. Interface design with a right modality and not give a cognitive load functions will ensure disabled learner more understand, receive and process information clearly.

This paper is structured as follows: In Section II, a discussion of the related works is given. Section III describes the methodology followed by results and discussion in Section IV and Section V conclude the study and suggestion for future work.

II. Related Work

The decision making process involves a competitive, complex and combative environment [7]. Everyone is a decision maker and making better decisions in life is a practice and skill that we need to consider and develop. Choosing the right criteria for a decision maker is one of the crucial aspects that needs to be analyzed. Accessing the problem and finding the solution through the available alternatives from multiple criteria is one of the multi-criteria decision making approaches (MCDM) [8]. Providing a solution with a variety of alternatives based on the right perspective and tools may help to cater to the problem at hand.

The variations of technology that are available with sophisticated attributes and dynamic would lead to various solutions for the disabled learner to choose the best assistive technology in terms of adaptability and availability [9]. Certain aspects and approaches would have issues and consideration which should be highlighted in the research based on the adaptability and suitability of technology to a disabled learner user. Based on Assistive Technology Industry Association, assistive technology can be described as a device, apps and principle that may help to enhance everyday life of disability person includes learning and working. [10]. Personalization is one of the main factors that needs to be considered in education as we are aware every learner has differences and this concern is recognized by educators and researchers [11].

Technology innovation is widely expanding in recent development where the gap between innovation and innovative products is reducing [12]. A variety of inventive products are available nowadays but how far does the product have an innovative value to the disabled learner? The attributes character of an innovation learning technology needs to be adapted and accepted by disabled learner users as it can reduce the burden that they carry. A study by Rogers state that the characteristics of innovation may influence the decision maker to adopt or reject the innovation [13].

Multi-criteria decision making (MCDM) has methods that are suitable to evaluate and make decisions by providing the best option of an alternative in order to choose the best criteria that make up the solutions [14]. Every piece of software has a uniqueness of attributes, and each of decision-makers have different point of views towards the attributes, hence the process of selecting the appropriate of software is a difficult task [15] .A variety of alternatives is provided which may optimize the solutions and cater to the problem that has arisen. The analytical hierarchy process (AHP) reflects on quantifying and evaluating intangible factors and choices in a multi objective environment and leads to the structuring of a complex structure in a decision making process [16].

A pairwise comparison between the attributes of criteria, sub- criteria and alternatives is the process to evaluate and optimize the solutions in the analytical hierarchy process (AHP). The analytical hierarchy process (AHP) is one of the decision making tools that can determine the importance of criteria and sub-criteria [17]. The formulation of an algorithm from the analytical hierarchy process (AHP) may help the teachers and parents of disabled learners to find the best design of learning technology that suits their children's ability and capacity.

III. Methodology

A survey was conducted in two (2) special schools under the Ministry of Education Malaysia (KPM), located in Johor, Malaysia. The survey was distributed among the school teachers who are the right people involved in the learning process of disabled learners. A total of 30 survey data were collected during the survey process. The main objective of the survey process is to collect feedback based on the suitability design of learning technology that is needed for disabled learners in a classroom learning environment.

A. Overview of AHP Method

In order to choose the best selection of design learning technology for disabled learners, the following method of flow analytical hierarchy process (AHP) should be considered. The first step in the analytical hierarchy process (AHP) is to define an objective based on the problem that has arisen in the situation. The first level represents an objective or goal, while the second level represents criteria, and the lowest level is alternatives. Once the hierarchy level is formed, the pairwise comparison is compared and analyzed based on each level of hierarchy. The details of process development of hierarchical model are per shown as Fig. 1 while procedure of flowchart of analytical hierarchy process are per shown as Fig. 2.

B. Development of Hierarchy Model

In this step, the hierarchy model of this study is developed. The three levels of hierarchy model in this study are described as the Fig. 2 as follow

Fig. 2.Design Learning Technology of Disabled Learner [1]

- Overview the situation
- Organize multiple criteria
- Asses multiple criteria
- Evaluate alternatives based on the assescriteria
- Rank alternative
- Comprise the judgment based multiple expert

Fig. 1. Process Development of Hierarchy Model.

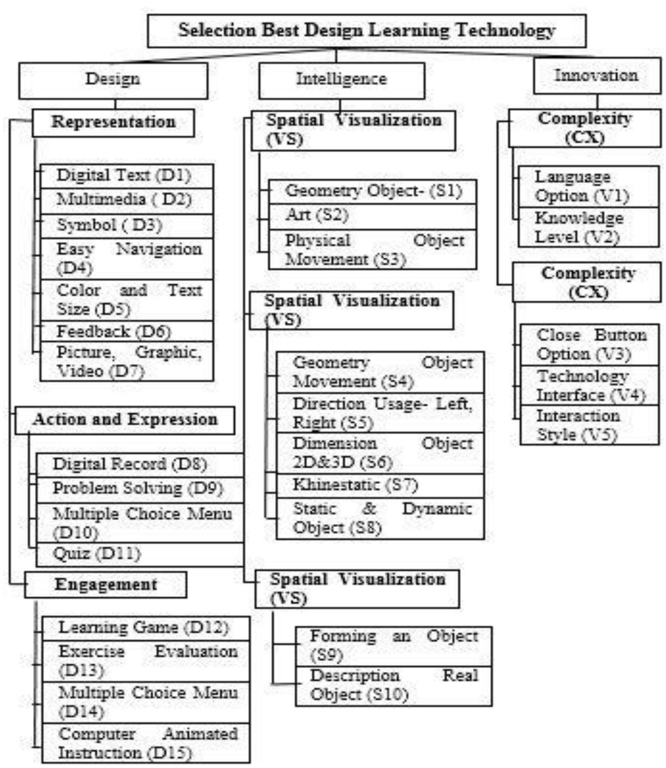


Fig. 2. Flowchart of Analytical Hierarchy Process (AHP) [1].

C. Steps in the Analytical Hierarchy Process (AHP)

The judgment matrix depends on the availability of the level hierarchy models. The following are the steps taken when applying an analytical hierarchy process (AHP):

a) Firstly, the structure of a hierarchy is based on setting goals that are made up of criteria, sub-criteria and alternatives.

b) Secondly, a pairwise comparison matrix for criteria, sub- criteria and alternatives is constructed as below:

$$A = \begin{bmatrix} a^{11} & a^{12} & a^{1n} \\ a^{21} & a^{22} & a^{2n} \\ a^{n1n2} & \ldots\ldots & a^{nn} \end{bmatrix} \tag{1}$$

Wherein $A_{ij,}$ i, j = 1,2……n, i= row, j=column, A|j=1 for ii=j

Based on the equation (1) above, the matrix *A* has three rows and three columns wherein a^{11} represents first row and first column.

A_{ij}=1 for ii=j

A_{ij}=$/\frac{1}{Aji}$ for i=j

c) Thirdly, the weight for criterion and alternatives selected from the matrix is calculated using the normalization procedure:

Calculation of data per row:

$$Wi = \sum_{j=1}^{n} aij, i = 1, 2, ...n \tag{2}$$

Wi in the equation (2) above represents the total of the sum of each column wherein *n* represent column and j is row = *i+j+n*

Value of weighted normalization:

$$Wi = \frac{\sum_{j=1}^{n} aij, i = 1, 2, ...n}{\sum_{k=1}^{n}\sum_{j=1}^{n} akj}, 1 = 1, 2....n \tag{3}$$

Equation 3 above shows on how to calculate a priority vector wherein $\sum_{j=1}^{n} aij$ represents the number of each column is divided by the total of each column that represents by $\sum_{k=1}^{n}\sum_{j=1}^{n} akj$

Eigen value:

$$Wi = \frac{1}{n} + (A1 + A2 + A3 + ... + An) \tag{4}$$

Eigenvalue is calculated by equation (4) wherein (*A1+A2+A3+…An)* are represented the total row is divided by the value of priority vector ($\frac{1}{n}$) and the results of value is formed by the equation (5). Hence, the vector weightage is multiplied by each number of columns and rows as per equation (6) to synthesize the vector weightage.

Vector weighting:

$$V = [V^1, V^2, ...V^n]^T \tag{5}$$

Finally the weight of all alternatives are synthesized with the weights obtained for each category as follows:

$$BxV = \begin{pmatrix} b^{11} & b^{12} & b^{1n} \\ b^{21} & b^{22} & b^{2n} \\ b^{n1} & b^{n2} & bm \end{pmatrix} \begin{pmatrix} V^1 \\ V^2 \\ V^n \end{pmatrix} \tag{6}$$

Next, the value obtained needs to do a consistency process to determine whether the judgment value obtained is consistent or not. The consistency formula is constructed as per shown in equation (7). After the judgment values obtained, the consistency index as stated as per equation (8) need to be done to determine whether the value is consistence or not, hence the percentage of value of consistency ratio can be obtained as per equation (9).

Maximum Eigen value:

$$\lambda \max = \frac{1}{n} = \sum_{i=1}^{n} \frac{(Aw)i}{wi} \tag{7}$$

Consistency Index:

$$CI = \frac{(\pi \max - n)}{n - 1} \tag{8}$$

Consistency Ratio: $\frac{\text{Index Consistency}}{\text{Average}}$

$$= \frac{(\lambda \max - n)}{r(n-1)} 100\% \tag{9}$$

Saaty states that the acceptable consistency ratio should be less than 0.10 and that even a ratio obtained less than 0.20 is also acceptable [18]. Evaluation of the consistency are generated by decision makers through judgment in pairwise comparison matrices in AHP [19].

IV. Result and Discussion

A. Perform Pairwise Comparisons for all Levels

The pairwise comparisons depend on the level of hierarchy of the elements. The pairwise comparisons in this study consist of criteria elements, sub-criteria elements and alternative elements. The evaluations of the pairwise comparisons for criteria are shown in Table I and the evaluations of the pairwise comparisons for sub-criteria elements are shown in Table II.

Table I shows the pairwise comparisons for all criteria that consist of learning design, intelligence, and innovation. The results show that design obtained the highest priority vector of 0.65, followed by intelligence (0.25) and innovation (0.10). Table II shows the pairwise comparisons of all sub-criteria that consist of representation, action and expression, engagement, spatial visualization, spatial relation, spatial orientation, complexity, and compatibility. The results show that representation obtained the highest priority vector of 0.57, followed by action and expression at 0.29, and engagement at 0.14 for the design sub-criteria. For the intelligence criteria, the highest alternative obtained from the pairwise comparison was spatial visualization at 0.61, followed by spatial relation at 0.27 and spatial orientation at 0.12. The complexity sub-criteria obtained a higher priority vector of 0.75, followed by compatibility at 0.25 from the results of the pairwise and synthesize comparisons.

B. Overall Priority Vector for all hierarchy

The following are the details of all combinations of priority vector, new vector, consistency index (CI), and consistency ratio (CR) obtained that consist of elements of criteria and sub-criteria, as shown in Table 3. If the value of consistency ratio (CR) is less than 0.1 ≤ 0.2, the evaluation is acceptable. Otherwise, if the value of consistency ratio (CR) obtained is more than 0.2, the judgment value needs to be revised by the decision maker until the value obtained is an acceptable value. Table III shows the results of all priority vectors for sub criteria that were obtained meanwhile Table IV shows the value of Random Index (RI) based on the number of judgments.

C. Develop Priority Ranking for All Hierarchy

The results of priority ranking for all hierarchies were developed based on the weightage value of the priority vector. To generate a final priority vector is by multiplying the priority vector of criteria by priority vector of each alternative of an objective [20]. Based on the pairwise and synthesizing matrix results, the overall priority vector for all hierarchies was obtained. The overall priority vector can be obtained by multiplying the priority vector of criteria with the priority vector of sub-criteria and alternatives. The calculation to obtain the overall priority vector is as follows:

(0.65) (0.57) = 0.37 wherein 0.65 (Priority Vector) for criteria and 0.57 (Priority Vector) for sub-criteria.

Table V shows the results obtained from the weightage of priority vector of all criteria and sub-criteria. In this study, the selection of best selection design is based on the results of the higher weightage of priority vector obtained from the synthesizing process. The best design in learning technology is needed to ensure that disabled learners can engage and enhance their performance in academic skills. The availability of technology present nowadays may not all be suitable for a user, especially disabled learners based on their ability and capacity. An innovative approach that integrates with a good and suitable design should be consider in development of learning technology of disabled learner.

TABLE. I. SYNTHESIZING MATRIX FOR ALL CRITERIA

Goal	**D**	**I**	**IV**	**PV**
Learning Design (D)	0.69	0.69	0.69	0.69
Intelligence (I)	0.17	0.17	0.17	0.17
Innovation (IV)	0.14	0.14	0.14	0.14

TABLE. II. SYNTHESIZING MATRIX FOR ALL SUB-CRITERIA

Goal	**D**	**I**	**IV**	**PV**
Representation (R)	0.60	0.69	0.43	0.57
Action and Expression (AE)	0.20	0.23	0.43	0.29
Engagement (E)	0.20	0.08	0.14	0.14
Spatial Visulization (SV)	0.60	0.69	0.43	0.61
Spatial Relation (SR)	0.20	0.23	0.43	0.27
Spatial Orientation (SO)	0.20	0.08	0.14	0.12
Complexity	0.60	0.69	0.43	0.75
Compatibility	0.20	0.23	0.43	0.25

TABLE. III. ALL PRIORITY VECTORS FOR SUB-CRITERIA [4]

Sub Criteria		Wt1	Wt2	C1
Learning Design (LD)				
R		0.65	0.57	0.37
AE		0.65	0.29	0.19
E		0.65	0.14	0.10
Intelligence (I)				
SV		0.25	0.61	0.15
SR		0.25	0.27	0.07
SO		0.25	0.12	0.03
Innovation (IV)				
CX		0.10	0.75	0.08
CT		0.10	0.25	0.03

TABLE. IV. RANDOM INDEX TABLE [5]

(n)	1	2	3	4	5	6	7	8
(RI)	0	0	0.58	0.9	1.12	1.24	1.32	1.41

TABLE. V. PRIORITY VECTORS FOR ALL CRITERIA AND SUB-CRITERIA [6]

Criteria	Priority Vector (PV)
D	0.65
I	0.25
IV	0.10
Sub Criteria	Priority Vector (PV)
R	0.57
AE	0.29
E	0.14
SV	0.61
SR	0.27
SO	0.12
CX	0.75
CT	0.25

TABLE. VI. PRIORITY VECTORS FOR ALL CRITERIA AND SUB-CRITERIA [6]

No	Criteria	Important of Element Interface	Weightage
1	Learning Design (LD)	Representation (R)	0.145
	Intelligence (I)	Spatial Visualization (SV)	0.044
	Innovation (IV)	Complexity (CX)	0.751

The integration of these three (3) criteria which are learning design, intelligence and innovation may increase and enhance the understanding of learning process of disabled learner. The integration of these criteria may attract the disabled learner to learn in interactive way in the class environment. The strength of disabled learner should be identified and embedded as one of an innovative approach in learning technology. As we've known that disabled learner have deficiency in cognitive level as compared with a normal learner. Hence, a special approach should be integrate with a usages of technology as it may help disabled learner more understands in their learning process.

Table VI shows the result of an importance of an element interface based on the selection of learning technology of disabled learners that consist of learning design, intelligence and innovation. The result shows that D1 has the highest value at 0.145 (representation), D8 with a value of 0.800 (action and expression), and D12 with a value of 0.800 (engagement). For the intelligence concept, the highest value is S1 with a value of 0.044 (spatial visualization), S4 with a value of 0.008 (spatial relation) and S9 (spatial orientation) with a value of 0.002. For the innovation concept, V1 has the highest value at 0.751 (complexity) and V3 with a value of 0.018 (compatibility).

D. Selection of Best Design for Learning Technology

The integration of these three concepts that are embedded as a special approach in designing learning technology can enhance the process of learning and attract disabled learners to learn in an interactive environment. The development of learning technology should be emphasized towards the needs of disabled learners based on their ability and capacity. Skills, strengths and abilities of disabled learners should be recognized to determine the best approach to overcome their limitations which at the same time may increase the quality of education for disabled learners.

V. CONCLUSION

The availability of learning technology nowadays can enhance and improve the learning process. Hence, the selection of best design of learning technology embedded with the right approach may attract and improve the learning process of disabled learners. The implementation of the analytical hierarchy process (AHP) in decision making may optimize the available alternatives to cater to the problem that has arisen. Based on the results obtained, it shows that learning design (LD) had the highest values obtained in the selection of design that is 0.145 followed by intelligence (I) 0.044 and complexity (CX) 0.751. Hence, these findings show the pilot results of selection determinants of interface criteria of learning technology for disabled learner that will use for developing learning technology of disabled learner. The best design is determined to help developers in developing learning technology for disabled learners that can be used in a classroom environment.

VI. FUTURE WORK

The authors have discuss on how to optimize the best selection design by using analytical hierarchy process (AHP) approached in selecting the best design of learning technology for disabled learner. But still there is a scope that is no covered yet by these findings hence future work will suggest a new hybrid method to implement and propose the evaluation these final findings of these study. A hybrid method will integrate and use by using different methods which is carried out for future work.

REFERENCES

[1] Altınay, F., Cagiltay, K., Jemni, M., &Altınay, Z, "Guest Editorial: Technology Support for Fostering Life-Long Learning of Learners with Disabilities," Educational Technology & Society, 2016, 19 (1), pp: 1–3.

[2] Pirani,Z and Sasikumar M."Assistive E-learning System for the Learning Disabled,"International Conference on Advanced Computing Technologies and Application, 2015, 45, pp:718-727.

[3] B.H. Benton-Borghi,"A Universally Designed for Learning (UDL) Infused Technological Pedagogical Content Knowledge (TPACK) Practitioners' Model Essential for Teacher Preparation in the 21st Century, J. Educational Computing Research, 2013,48(2), pp:245-265.

[4] Nasir, M.N.A and Efendi A.N.A.E." Special education for children with disabilities in Malaysia: Progress and obstacles Muhamad Nadhir Abdul Nasir," Geografia-Malaysian Journal of Society and Space 12(10).

[5] FernáNdez-LóPez, Á., RodríGuez-FóRtiz, M. J., RodríGuez-Almendros, M. L., & MartíNez-Segura, M. J. "Mobile learning technology based on iOS devices to support students with special education needs." Computers & Education, 2013, 61, 77-90.

[6] Zhuhadar, L., Carson, B., Daday, J., Thrasher, E., & Nasraoui, O. "Computer-assisted learning based on universal design, multimodal presentation and textual linkage", Journal of the Knowledge Economy,2016. 7(2), 373-387.

[7] Thirumalai, C., & Senthilkumar, M. "An assessment framework of intuitionistic fuzzy network for C2B decision making. In 2017 4thInternational Conference on Electronics and Communication Systems (ICECS).2017.pp. 164-167.

[8] Wu, H. M.,"Online individualised tutor for improving mathematics learning: a cognitive diagnostic model approach".2018.Educational Psychology, 1-15.

[9] Adebisi, R. O., Liman, N. A., & Longpoe, P. K.," Using Assistive Technology in Teaching Children with Learning Disabilities in the 21st Century".2015. Journal of Education and Practice, 6(24), 14-20.

[10] Taylor, H. (2019). Assistive Technology for Students with Learning Disabilities: Kurzweil 3000. Technology and the Curriculum: Summer 2019.

[11] El Mawas, N., Tal, I., Moldovan, A. N., Bogusevschi, D., Andrews, J., Muntean, G. M., & Muntean, C. H., "Improving STEM Learning Experience in Primary School by Using NEWTON Project Innovative Technologies". 2018. InInternational Conference on Computer Supported Education. Springer, Cham pp. 214-230.

[12] Wani, TA, Ali, SW, "Innovation Diffusion Theory: Review & Scope in the Study of Adoption of Smartphones in India," Journal of General Management Research.2015. 3 (2), 101-118.

[13] Rogers, E.M.Diffusion of Innovations, 5th ed.; 2003. The Free Press: New York, NY, USA.

[14] Aziz, N. F., Sorooshian, S., & Mahmud, F., "MCDM-AHP method in decision makings". ARPN Journal of Engineering and Applied Sciences, 2016.11(11), 7217-7220.

[15] Abdullateef, B. N., Elias, N. F., Mohamed, H., Zaidan, A. A. & Zaidan, B. B. 2016. An evaluation and selection problems of OSS - LMS packages. SpringerPlus. doi:10.1186/s40064-016-1828-y.

[16] Ahmad, Mohd Nazri, N. A. Maidin, M. H. A. Rahman a, "Conceptual Design Selection of Manual Wheelchair for Elderly by Analytical Hierarchy Process (AHP) Method: A Case Study, "International Journal of Applied Engineering Research,2017 vol: 12 (17) pp: 6710-6719.

[17] Cheng, E. W., & Li, H.,"Analytic hierarchy process: an approach to determine measures for business performance".2001. Measuring business excellence, 5(3), 30-37.

[18] Wedley, W. C., "Consistency prediction for incomplete AHP matrices. Mathematical and Computer Modelling, 199317(4-5), pp:151-161.

[19] Aguar, J. & Tur, A. 2019. AHP-Group Decision Making Based on Consistency (Ccm). doi:10.3390/math7030242.

[20] Simon, J., Adamu, A., Abdulkadir, A. & Henry, A. S. 2019. Analytical Hierarchy Process (AHP) Model for Prioritizing Alternative Strategies for Malaria Control 5(1): 1–8. doi:10.9734/AJPAS/2019/v5i130124.

Predicting IoT Service Adoption towards Smart Mobility in Malaysia: SEM-Neural Hybrid Pilot Study

Waqas Ahmed[1], Sheikh Muhamad Hizam[2], Ilham Sentosa[3], Habiba Akter[4], Eiad Yafi[5], Jawad Ali[6]
UniKL Business School (UBIS), Universiti Kuala Lumpur, Kuala Lumpur, Malaysia[1, 2, 3, 4]
Malaysian Institute of Information Technology (MIIT), Universiti Kuala Lumpur, Kuala Lumpur, Malaysia[5, 6]

***Abstract*—Smart city is synchronized with digital environment and its transportation system is vitalized with RFID sensors, Internet of Things (IoT) and Artificial Intelligence. However, without user's behavioral assessment of technology, the ultimate usefulness of smart mobility cannot be achieved. This paper aims to formulate the research framework for prediction of antecedents of smart mobility by using SEM-Neural hybrid approach towards preliminary data analysis. This research undertook smart mobility service adoption in Malaysia as study perspective and applied the Technology Acceptance Model (TAM) as theoretical basis. An extended TAM model was hypothesized with five external factors (digital dexterity, IoT service quality, intrusiveness concerns, social electronic word of mouth and subjective norm). The data was collected through a pilot survey in Klang Valley, Malaysia. Then responses were analyzed for reliability, validity and accuracy of model. Finally, the causal relationship was explained by Structural Equation Modeling (SEM) and Artificial Neural Networking (ANN). The paper will share better understanding of road technology acceptance to all stakeholders to refine, revise and update their policies. The proposed framework will suggest a broader approach to investigate individual-level technology acceptance.**

Keywords*—Smart Mobility; Internet of Things (IoT); Radio-Frequency Identification (RFID); Neural Networks; Technology Acceptance Model (TAM)

I. Introduction

In today's world, the term "Smart" depicts the intelligence and self-learning capabilities of non-living objects around human. The tangling of word "smart" as the smart-phone has, now, been diffused to many segments of life like smart-tv, smart-school, smart-home, smart-car, and smart-city etc. This smartness of the things is backed by numerous technologies i.e., RFID sensors, Internet of Things (IoT), Big Data Analytics and Machine Learning. The main purpose of smart objects is to provide the efficient and convenient way of living without human interference. Likewise, smart city manages the assets, resources, and services of urban areas to well-plan the dwelling, working, and commuting for inhabitants. Smart mobility is one of the main elements that enhances the smart city management by ensuring the safe, clean, and economical commuting and transportation mechanism. Smart mobility aims to cope up with numerous challenges such as traffic management, congestion mitigation, environmental impact control and infrastructure safety etc. through digital tools and techniques.

The urbanization around the world has reached to 55% while Malaysian urban density is expanded to 76% which is expected to surge at 82% in next 10 years[1]. Denizens of urban Malaysia have higher preference of purchasing car in cash on hand state [2]. Therefore Malaysian have become the 3rd highest level car ownership nation worldwide, where 93% of households own and utilize at least one vehicle [3]. The aggregate number of registered automobile is 88% (28.2 M) of total population (32.4 M) [4]. It is followed in excess use of personal vehicle over public transportation that concluded in challenges of higher congestion level, environmental impact, road safety, infrastructure impairment[2]. Due to this reason, the commuters spend more time on roads and suffer the traffic jams with higher pace as compared to the preceding years. Additionally, a survey by The Boston Consulting Group, indicated that inhabitants of Kuala Lumpur stuck in road congestion around 53 minutes daily and looking for parking spots takes their 25 minutes [5]. The situation becomes aggravate at toll plazas, particularly in vicinity of urban areas, where long queues waste the productive time and fuel, damage the road, and contaminate air quality. A smart mobility study in Klang Valley explored the alarming situation of damages due to outnumbered vehicles on roads by estimating productivity loss of RM 5 billion per annum [6]. As per World Bank report, the economic losses due to congestion across Klang Valley in 2014 were RM 20 billion, around RM 52 million daily. While stuck in traffic means being non-effective citizens, who waste the total time with the value of RM 10 billion to RM 20 billion annually by doing nothing. Similarly, the wasted fuel due to congestion surges around RM 2 billion. This excessive fuel burning poses the great environment danger and social threat. The overall price of traffic congestion in Klang Valley was projected at 1.1 to 2.2% of GDP in 2014 [7]. Being rolling down slowly on roads results in impairment of infrastructure that also costs the government. In urban areas of Malaysia, the vehicle average speed has been lowered down due to congestion that demands authorities to keep the facilities up to date [8].

The authorities have persuaded various ways to tackle such issues by penetrating the public transport network extension and ride sharing e-hailing services etc. To manage the private cars on roads the use of RFID sensor along E-Wallet has launched in 2019. This RFID tag is initially utilized for paying the e-toll or electronic toll collection (ETC) on toll plaza which is being implemented for smart parking and electronic road pricing (ERP). The RFID Tag is affixed on windscreen or headlamp of vehicle and then it is linked with E-wallet account. Previously, Intelligent Transport System (ITS) of Malaysia had implemented the infrared based On-Board-Unit (OBU) technology, SmartTag for e-tolling services in 1998. However due to lower adoption level i.e., 28% of registered vehicles, it was discontinued in August 2018. The new RFID sensor system aims to function through E-Wallet app for paying the tolls, parking fee and congestion price. It will eradicate the lag time in long queues, operates the traffic flow, manage the parking system through smart phone. Electronic road pricing (ERP) or congestion pricing will regulate the traffic by charging the vehicles during peak time. ERP will also regulate the traffic around schools, hospitals, and parks. By implementing the RFID sensors, ITS Malaysia will instigate Multilane Free Flow (MLFF) for gate-less or open-road tolling on highways. Embedding smart mobility services like RFID sensors to metropolis and enhancing its acceptance is the venture that will yield the benefits to government, citizens, and environment.

This IoT based RFID service has multiple benefits for users and government [9] but it has certain complexities that might hinder the proper implementation process. The RFID tag uses the E-Wallet payment method which requires users' personal information, bank account, debit/credit card details in order to complete the transaction. While cashless payment method in Malaysia depicted that Mobile wallet is the least used method as only 8% citizen using E-Wallet [10]. There are various studies, comprised of the technicalities of IoT enabled ETC system that work through RFID sensors, have suggested the on-going improvement and implications[11]–[14]. But there are very few studies focusing on personal and social factors of using this IoT based technology[15], [16]. Besides this, the user-based studies pertaining the motives and aspects that were backing the lack of acceptance and affected the lesser usage of SmartTag by motorists are also largely unknown. The literature on users' behavior towards transportation technology is limited in Malaysia while there is a gap in assessing mobile wallet usage for smart mobility services.

Acceptance and adoption of new technology always depend on perception and behavior of user towards the system. In era of technology advancement when Industry 4.0, Cloud computing and IoT are infused around, the concept of smart mobility is of higher priority [17]. Acceptance and continuation of innovative system such as RFID service along Mobile Wallet, is a personal choice that is mainly linked with human attitude and behavior [16]. Adopting the digital technology requires the innovative personality and digital dexterity that comprises of ability to understand the technology and passion to get benefit from it in daily activities. The main factors that play important role in users adoption are based on comfort, convenience and usefulness of technology [18]. Perceived enjoyment, trust, perceived behavior control positively impact the behavioral intention of user to utilize the ETC system[16]. As there are various service quality assessment model (SERVQUAL) in literature [19], the quality assessment of IoT based services also craves for the updated SERVQUAL model dimensions [20]. In electronic services, the matter of service quality always possesses the main factor towards customer trust, loyalty and branding of service. Personal factors like level of technological understanding and security concerns of using technology are also vigorous drivers of adopting the technology [21]. Using the Mobile Wallet service has shown the various issues of customer privacy and reluctance [22] while users privacy concerns for digital wallet are validated by extending Technology Acceptance Model (TAM) [23]. In scenario of IoT service acceptance, the intrusiveness concerns is verily undertaken in previous studies [24]. On the other hand, innovative service penetration is achieved by the tech-awareness capability of users [25]. Measuring the level of technology acceptance in various studies involved the personal understanding and usage of technology [26], [27] that describes as the digital dexterity. While social media role in disseminating the positive or negative information also intrigues the individual's behavior towards the technology use [28]. Society always backing up the human behavior in order to accept or reject the technology [29]. Society views are disseminated physically or virtually i.e., internet, both have impact on behavior to decide for technology acceptance [30]. Such antecedents of system use can better represent through research model or theoretical framework to understand the relationship towards users' adoption behavior.

The study proposes an adoption model i.e., IoT-TAM model to predict the attitude and behavior of motorist towards acceptance of Internet of Things based smart mobility service. This model aims to answer about the contextual predictors that influence the smart mobility service adoption. It will elaborate the support of existing theories in smart mobility. The model will also explore the impact of personal characteristics on behavior to adopt and use the smart mobility services from digital environment perspective. The theoretical framework will also depict the service quality model for IoT platform. The impact of circumstantial variables (i.e., digital dexterity, intrusiveness concerns, social electronic word of mouth and social norms) together with two theories in digital context (i.e., TAM and SERVQUAL) will instigate initial step towards filling the gap in determining the constructs of the smart mobility service adoption. The proposed model will be fully corroborated through a widespread data collection across Malaysian cities, which nurtures through the pilot study testified here. The outcomes will eventually support in illuminating the vital factors in adoption of smart mobility services in Malaysia.

II. Literature Review

A. Theory of Technology Acceptance Model

The research model adopted in this study is rooted in Technology Acceptance Model (TAM). TAM was proposed as a basic framework for tracing the impact of external factors on internal beliefs, attitudes and intentions of human[31]. TAM also projected that human attitude towards information systems is a consequence of two major principles as portrayed in Fig. 1. These include perceived usefulness (PU), that the concepts as an individual's work efficiency will improve by utilizing certain technology. When system facilitates work efficiency, the user apprehends the system in positive way (attitude). This optimistic attitude increases the eagerness to involve the system usability. While ease of use i.e., PEOU refers to the level of convenience in utilizing the technology. Individual perceives the PEOU as feasibility element to use the system. And these both construct effect the individual perception and form the antecedent attitude, refers as evaluative impact [32]. These both constructs build the positive or negative attitude towards intention to accept the technology. These two dimensions are nurtured by individual's exposure to external factors that are correlated with specifications of information system and the environment. TAM model directed that usage of information system is managed by behavioral intention that mutually impacted by the attitude of individual and technology usefulness [31]. However, people express the intention to accept the information system upon knowing its usefulness despite the despite the attitude towards the system. [31].

Many studies have justified Perceived usefulness as most influencing construct in TAM model as it defines the user perception towards the importance of information technology use [33]. In transportation field, perceived usefulness possesses the considerable influence on the attitude and behavior of technology users. As user acceptance behavior of electronic vehicle is supported by perceived usefulness [34]. Its prominence in effecting the behavior intention towards the adoption of vehicle navigation system is also acclaimed [35]. It has significant impact on user's attitude [36] and intention towards acceptance of e-tolling system [37]. This would allow the researchers to examine if, in smart mobility service adoption context, perceived usefulness would have any significant influence on attitude and behavioral intention. This is hypothesized as:

H1: "Perceived Usefulness (PU)" will positively effect "Attitude (ATT)" to adopt smart mobility services.

Perceived ease of use (PEOU) is described as "the degree to which a person believes that using a particular system would be free of effort" [38]. It is also the powerful determinate in shaping the user attitude by increasing the system usefulness that eventually effect the behavioral intention of the system [31]. Motorists attitude towards the usage of transport technology is backed by the perceived ease of use [39] while using the electronic tolling service attitude mostly affected by this factor [36], [37], [40]. The more convenient and easiness motorists gain from using IoT service on roads, the more satisfied they are, and therefore, the more likely they are eager to adopt the services. In perspective of this study, the author hypothesizes that perceived ease of use (PEOU) will have a positive influence on the perceived usefulness (PU) and attitude to use the IoT System. This would allow the researcher to examine if, in this context, perceived ease of use would have any significant influence. This is hypothesized as.

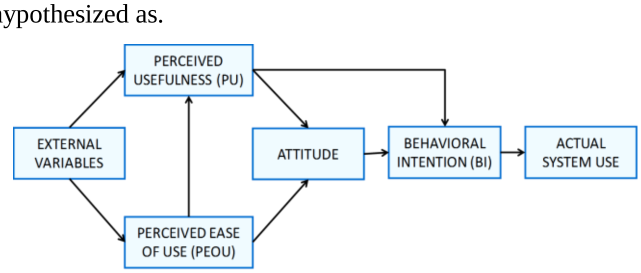


Fig. 1. Technology Acceptance Model.

H2: "PEOU" will positively effect "Attitude" to adopt smart mobility services.

H3: "PEOU" will positively effect "Perceived Usefulness" to adopt smart mobility services.

In technology adoption studies, attitude indicates positive or negative feelings of users towards system usage behavior. Theory of Planned Behavior foresees that the more fortunate an individual assesses the certain behavior, more probably he will be anticipated to perform that behavior [32]. In transportation sector it has significant relationship with behavioral intention [41] and specifically for ETC acceptance [42]. As motorists get more positive felling from using IoT service on roads, the more inclined they are, and therefore, the more likely they are eager to adopt the services. This would allow the researcher to examine if, in this context, attitude would have any significant influence on behavioral intention. This is hypothesized as:

H4: "Attitude (ATT)" will positively effect "Behavioral Intention (BI)" to adopt smart mobility services.

B. Service Quality

While dealing with electronic services, there is significant relationship among various users' related factors such as service quality, perceived usefulness (PU), customer's satisfaction and permeance of system usage. Service usefulness is vitally managed by service quality [43] that determines the users' behavior. In transportation research, user perception and awareness towards the service quality are of substantial importance [44]. Service quality for IoT based services can be assessed through service quality models measuring the various technological services like SERVQUAL [45], E-SERVQUAL [46] and SSQUAL [19] models.

SERVQUAL model was theorized to establish the scale of quality of services provided by organization and governments. [45] abstracted service quality in the five measurements concepts comprise (1) Reliability, (2) Responsiveness, (3) Assurance, (4) Empathy, and (5) Tangibility. This model had various extension in field of modern day technology like E-SERVQUAL, WebQual, SiteQual, IRSQ, eTailQ, PeSQ, SSTQ [19]. Later, as businesses shifted to digital services for their customer and clients through websites or web portals, the aspects of services quality updated to level of privacy of clients' data on websites, system availability, efficiency of websites and fulfilment the purpose of service. This model called as E-SERVQUAL or E-SQ [46]. It has widely used in measuring the level of service quality for websites and mobile application. An online support system for employees by government in Spain is assessed by scales of service quality (i.e., E-SERVQUAL) [47]. In Taiwan, group-purchase criteria from social media website like Facebook assessed through key quality characteristics of E-SERVQUAL [48]. Increasing labor costs have invigorated businesses to explore additional self-service alternatives that let customers to carry out services for themselves. Information technology permitted companies to practice a variety of self-service technologies (SSTs) that rise customer partaking. Self-Service Technology Quality model (SSTQUAL) was introduced by [46] that consist of 07 scales quality measures as 1) Functionality, 2) Enjoyment, 3) Security/Privacy, 4) Assurance 5) Design, 6) Convenience, 7)

Customization. However, to assess the quality of IoT based services, 04 dimension-based service quality measure is proposed that consists of privacy dimension from E_SERVQUAL and SSTQUAL, Functionality from SSTQUAL, Efficiency from SERVQUAL and Tangibility from SERVQUAL.

Several dimensions of information system (IS) service-quality have been established to asses IT success. System reliability, functionality, ease of use, system accuracy, response time turnaround time, completeness, system, flexibility, reliability, assurance and security are the various acknowledged dimension. SERVQUAL is most famous and its extensions for online or electronic business like E-SERVQUAL, SiteQual, WebQual etc. are commonly used in assessing the service quality in perspective fields [45], [46]. While currently technologies like Internet of Things, Big Data, Artificial Intelligence etc. go for the scale pertaining the measures of privacy/security, efficiency, functionality and tangibility. Service quality positively influences the constructs of TAM model [49]. Service quality model has positive impact on TAM variables in various sectors of tech-enabled services[50], [51].

In perspective of this study, the authors hypothesize that IoT-service quality will have a positive impact on the perceived usefulness, perceived ease of use, and behavioral intentions to use the IoT based RFID System. This is hypothesized as:

H5: IoT Service Quality (IoT-SQ) will positively effect "Behavioral Intention (BI)" to adopt smart mobility services.

H6: IoT Service Quality (IoT-SQ) will positively effect "Perceived Usefulness (PU)" to adopt smart mobility services.

H7: IoT Service Quality (IoT-SQ) will positively effect "Perceived Ease of Use (PEOU)" to adopt smart mobility services.

C. Intrusiveness Concerns

Intrusiveness concerns based on consumer's view that the service provider indecently intrudes into his or her personal life. It is assumed that when RFID technology is intricate, the privacy couldn't be absolutely guarantee, however it's interloped upon at multiple levels [52]. Perceived intrusiveness can be an obstacle to the use of custom digital services. RFID being a prevalent and ubiquitous innovation has prompted many debates because of privacy apprehensions. Towards adoption of the IoT service, intrusiveness concerns is one of the basic challenge [53]. Upon installation of app, user allows to share the a lot of information that might go for intrusiveness [54]. While intrusiveness has found significance towards the TAM variables [55]. In particular, the researchers assume that intrusiveness concerns of users will negatively impact the TAM variables i.e., perceived usefulness, perceived ease of use, and behavioral intention toward RFID tag acceptance. These hypotheses are formulated to evaluate intrusiveness as an exterior factor variable as well as to prove empirical relations with TAM model.

H8: Intrusiveness Concerns (IC) will negatively effect "Perceived Usefulness (PU)" to adopt smart mobility services.

H9: Intrusiveness Concerns (IC) will negatively effect "Perceived Ease of Use (PEOU)" to adopt smart mobility services.

H10: Intrusiveness Concerns (IC) will negatively effect "Behavioral Intention (BI)" to adopt smart mobility services.

D. Digital Dexterity (DD)

An individual's alertness or vigilance towards evaluating the innovation is personal innovativeness (PI). PI is the extent to which an individual perceives inclination and tendency to adopt and involve in usage of novel innovations. High level of personal innovativeness in users is projected to strengthen their positive intent towards the new innovative system or process. This can be used to classify early adopters who can either work as change mediators or else be directed specifically for adoption when resources are narrowed. To conclude, this concept of personally inclined behavior towards innovation is possibly considered to enhance broad concentrated models of IT implementation that comprised of paradigms other than individual beliefs or observations towards technology embracing decisions[26]. The term has been updated for digital environment context and this study will amend this as Digital Dexterity (DD).

The behavioral sciences with individual psychology, though, proposed that collective manipulativeness and impact and individual mannerisms for instance personal innovative behavior are possibly vital factors of technology acceptance and more imperative component in prospective users' assessment to use [56]. [57] said that individuals with higher commitment towards the innovation is backing the very success of the innovation process. The construct determines the innovative behavior of an individual in a scale from high to low, hence assisting to recognize individuals who are ascent to accept innovations before or after than others [58]. In study of an information technology associated to new innovations, it is essential to discover individual personal innovativeness towards the innovation. It is an important predictor or anticipator towards technology acceptance [59]. While in transportation studies, it also showed significance and positive relationship with TAM variables like perceived ease of use, perceived usefulness, attitude and behavioral intention to make use of the technology [60]. In accordance with the above, this study contends that drivers with higher level of innovativeness will positively inclined towards the RFID tag, therefore we hypothesis that:

H11: Digital Dexterity (DD) will positively effect "Perceived Ease of Use (PEOU)" to adopt smart mobility services.

H12: Digital Dexterity (DD) will positively effect "Perceived Usefulness (PU)" to adopt smart mobility services.

H13: Digital Dexterity (DD) will positively effect "Behavioral Intention (BI)" to adopt smart mobility services.

E. Social Electronic Word of Mouth (Social-eWoM)

Social electronic word of mouth is an interaction among online users by social interactive websites has proved to be the most preferred medium of electronic word of mouth (eWOM) system. [61]. Sharing and collaborating with consumers via social networking websites such as Facebook, twitter, etc. facilitate customers with probably unbiased information of products on anonymity base [62]. Electronic word of mouth has substantial effect on traveling sector. It is considered to be the valuable means of information influencing the travelers' behavior regarding their plans [63]. It is discovered that users frequently get online reviews submitted or shared by other tourists are conversant, pleasant, and trustworthy than communicated by journey facilitator organization [64]. Importance of dispersion of information through electronic mean for travelling distinguished by [65]. This eWoM proved significant impact on TAM variables in acceptance of ETC technology [40]. As drivers know more about this IoT based technology over internet they will tend to experience this. In this study perspective, it is hypothesized that social EWOM will have a positive impact on perceived usefulness, perceived ease of use, attitude and behavioral intentions to use the IoT based RFID ETC System. This is hypothesized as:

H14: "Social Electronic Word of Mouth (SWoM)" will positively effect "Intention (BI)" to adopt smart mobility services.

F. Subjective Norm (SN)

Higher the realization from other people's influence i.e., subjective norms, the more inclination by individual's attitude towards performance. [66]. Researchers depicted that more socially influenced individuals are likely to adopt certain technology. Subjective norms proved to be one of the key factors with extended TAM in numerous papers strengthening the user behavior towards the studied system and found positive and significant relationship with TAM variables [35]. In same way, society also firmed its impact on individual in transportation technology acceptance studies. Studies pointed out the social influence role in acceptance of E-tolling technologies in Taiwan and Indonesia [57].

The more socially influenced opinions motorists get for using IoT service on roads, the more indulged they are towards the technology, and therefore, the more likely they are eager to adopt the services. The authors hypothesize that subjective norm will have a positive impact on behavioral intention to utilize the IoT based RFID System. This would allow the researchers to examine if, in this context, subjective norm would have any significant influence. This is hypothesized as:

H15: "Social Norm (SN)" will positively effect "Intention (BI)" to adopt smart mobility services.

G. Theoretical Framework

A Theoretical framework is established after evaluating the well-known models in the domain of technology adoption. The model has origin in Technology Acceptance Model, and it is extended with new concepts. The proposed research framework comprises of five independent variables i.e., digital dexterity, perceived intrusiveness, service quality for Internet of Things services, social electronic word of mouth and subjective norm while there are four dependent variables i.e., perceived ease of use, perceived usefulness, attitude and behavioral intention. The research model is portrayed in Fig. 2, the shaded area representing the classical Technology acceptance model (TAM) variables.

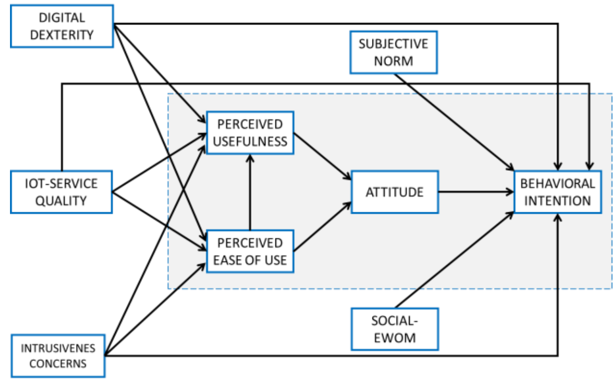


Fig. 2. Proposed Theoretical Framework (IoT-TAM).

III. Methodology

This study is based on assessment of digital factors' impact on human behavior therefore philosophy of the research is positivism. In positivism, the causal research is undertaken which deals with the cause and effect analysis by using the quantitative data. The research approach of this quantitative paper is deductive that deals with the adoption of established theory and then extends it as per the study requirement. TAM and SERVQUAL theories are used and extended in smart mobility context. Research strategy of study is Survey by which the data has collected through mono-method i.e., quantitative data technique with cross-sectional approach of time horizon [67].

Primary data for this pilot study was collected through survey. A research questionnaire was adopted from previous studies and developed for the smart mobility context. Questionnaire instruments for technology acceptance model (TAM) were taken from [38], [68] and service quality items were adopted from [19], [45], [46]. Other factors instruments were adopted as digital dexterity [26], [69], intrusiveness concerns [55], social electronic word of mouth [70] and subjective norm [71]. The questionnaire was divided into 03 section. First section focused on demographic variables, section 02 pertained the variables questions and third section consisted of respondents view on RFID system utility. The section 02 was based on measures smart mobility adoption factors such as digital dexterity, Social EWOM, Service quality, perceived usefulness, perceived ease of use, attitude, and subjective norm. To assess the response towards acceptance of smart mobility service, a five-point Likert-scale was utilized. The scale scored as range from 1 = "strongly disagree" to 5 = "strongly agree".

The analysis for pilot study mainly focuses the validity and reliability of questionnaire instruments which paves the way for full study data analysis. However, this paper has undertaken the SEM-Neural analysis approach for pilot data. To analyze the reliability of questionnaire items, Cronbach alpha, was measured through SPSS 25. Composite reliability, factor loading, discriminant validity, HTMT and r-squared values were analyzed through SmartPLS tool. Structural modeling of proposed framework was also measured by SEM in SmartPLS. Furthermore, Normalized importance of predictors towards the dependent variable (intention to adopt RFID services) were measured by Neural Network in SPSS25.

IV. Results

A. Descriptive Results

To validate the research model, a pilot survey was performed in Klang valley. The terms Klang Valley is used for the geographical location based on vicinity of Kuala Lumpur district, Selangor District, Federal territories, and surrounding areas. It is also known as Greater Kuala Lumpur and its population is 7.8 million which is 25% of country population [72]. A field survey was organized in this region to collect the response from drivers. As initially RFID service is used for paying toll on highways therefore rest areas on highways were considered for data collection point as illustrated in Fig. 3. Three points on highways were taken for the survey as depicted in Table I.

The survey process was completed in 04 weeks and responses were collected through face to face survey. The questionnaires were printed in both English and Malay language. The respondents were introduced with the purpose of study prior to getting their views. Total 80 responses were collected for this pilot study which are suitable for required analysis. The description of respondents according to section 1 of questionnaire i.e., demographic variables are described in Table II. The particulars of demographic data showed the more participation of male, private sector employees, age group 31 to 40, monthly income RM5000 to RM8000, single car ownership and daily toll use frequency.

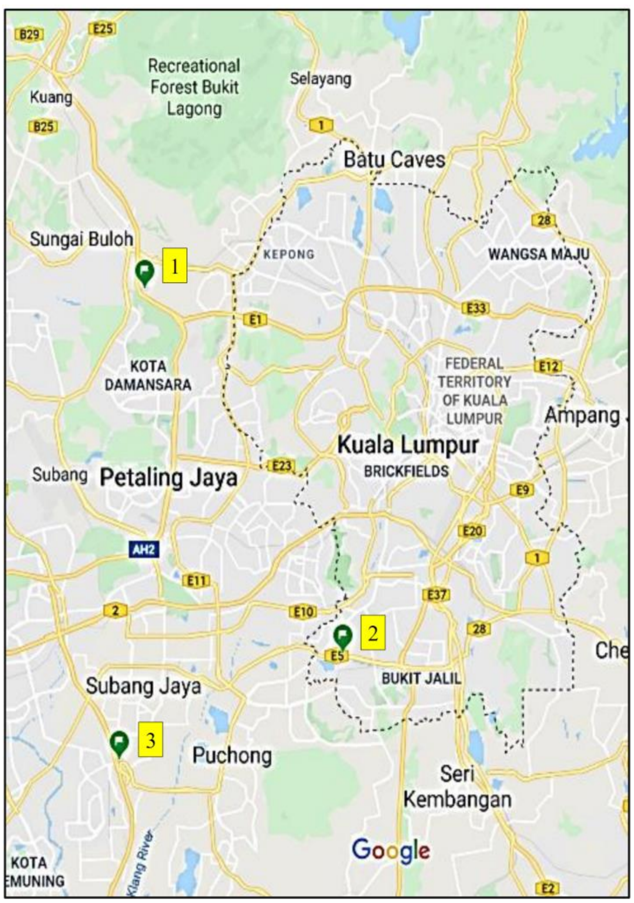


Fig. 3. Survey Location Point (Klang Valley).

TABLE. I. Survey Points

Point	Location Name
1	OBR - Sungai Buloh (PLUS Expressways)
2	R&R - Awan Besar (KESAS Highway)
3	OBR - USJ21, Subang Jaya (ELITE Highway)

TABLE. II. Demographic Results

Gender	%	Occupation	%
Male Female	84 16	Govt Sector Private Sector Student Business Others	10 51 14 20 5
Age	**%**	**Income Level**	**%**
18-30 31-40 41-50 >51	28 42 21 9	Below RM 2,000 RM 2,000 – RM 5,000 RM 5,001 – RM 8,000 RM 8,001 – RM 15,000 More Than RM 15,000	4 31 48 16 1
Race	**%**	**Education**	**%**
Malay Chinese Indian	74 15 11	Diploma/Certificate Primary School Secondary School University Others	6 6 33 54 1
Vehicle Ownership	**%**	**Toll Use Frequency**	**%**
0 1 2 3	3 62 29 6	Daily Weekly Monthly	76 20 4

B. Preliminary PLS-SEM

To assess the instrument validity and reliability of measuring in such way that the questionnaire items representing the variables are measuring the characteristics of variables. The tests of reliability to assess internal consistency (i.e., Cronbach alpha reliability and composite reliability) and validity (construct validity, discriminant validity) are performed. For this purpose, numerous analysis i.e., Cronbach Alpha Reliability test, Composite Reliability, Average Variance Extracted (AVE), Discriminant Validity HTMT Test were performed on SmartPLS. The recommended value of Cronbach Alpha is classified as excellent when higher than 0.9, strong when higher than 0.8 and acceptable when higher than 0.7. In case of its value less than o.7, the instruments have to recheck for exclusion of non-reliable items. Similarly, composite reliability is another internal measure of items consistency and its threshold value is 0.7 which is achieved by this pilot study analysis. These reliability tests will ensure the model is fit for causal analysis. In Table III, all the constructs are fulfilling the recommended criteria of reliability test as Cronbach alpha value of all variables is higher than 0.8 and some constructs have more than 0.9. It means the model variable are internally consistent for further analysis [73]. In next step, convergent validity was measured by average variance extracted (AVE) and its minimum accepted value was 0.5. The value AVE in Table III, shows all constructs are valid. However, Service quality has the value less than 0.5 because it is second order construct. While validity is the scale used for only first order construct measurement. Validity tests ensures there are no corelated variables in the model who are demonstrating the same characteristics. Discriminant validity is one the most important test for guaranteeing the success of pilot study. There are numerous methods used for discriminant validity like Fornell-Lacker Criterios, HTMT etc. This paper has undertaken the HTMT (Heterotrait-Monotrait Ratio of Correlations Test). The general rule of correlation among model factors that the value between two variables should be less than 0.9, the same rule is applied in HTMT measure. In Table IV, the discriminant validity values are accomplishing the set standards. There are no highly co-related variables in the construct set that needs to exclude from the model.

The proceeding analysis of pilot survey was based on outer loading. In this phenomenon, the contribution of items or questions towards their respective construct is measured. The standard acceptance value for outer loading is more than 0.7. In Table V, indicators or items of respective constructs are shown by their respective outer loadings. Here some items had not met the criteria i.e., value should be greater than 0.70 [74]. Therefore, items from Perceived ease of use (PE5), Digital dexterity (DD4), Social electronic word mouth (WM3, WM4, WM5) were deleted from model due to the outer loading value less than 0.70. In Table V, outer loading values are normal and showing the accepted impact level.

C. SEM-Neural Analysis

Towards causal structuring of model, preliminary analysis was conducted internally between the construct and respective items. The next step of analysis was to assess the causal relationship between the model constructs therefore Structural Equation Modeling (SEM) and Artificial Neural Networking (ANN) were employed. However, for pilot study with small sample size the SEM and ANN are not highly validated in terms of assessing detailed impact of relationships, data normality, RMSE etc. This pilot survey aims to pave the way for full scale study. But these tests can well-predict the behavior through overall variance in SEM and normalized importance in ANN while neglecting the sample size[75], [76].

TABLE. III. RELIABILITY AND VALIDITY

Variables	Items	Cronbach Alpha	rho_A	Composite Reliability	AVE
Attitude (ATT)	04	0.938	0.939	0.956	0.844
Behavioral Intention (BI)	05	0.964	0.964	0.972	0.874
Digital Dexterity (DD)	06	0.894	1.022	0.907	0.625
Efficiency (EF)	04	0.897	0.898	0.928	0.763
Functionality (FN)	04	0.822	0.828	0.882	0.653
Intrusiveness Concerns (IC)	04	0.928	0.932	0.949	0.824
IoT-Service Quality (SQ)	17	0.924	0.929	0.934	0.451
Perceived Ease of Use (PEOU)	04	0.933	0.934	0.952	0.832
Perceived Usefulness (PU)	05	0.957	0.959	0.967	0.855
Privacy (PR)	05	0.897	0.898	0.924	0.711
Social EWOM (SWOM)	03	0.865	1.152	0.879	0.551
Subjective Norm (SN)	04	0.911	0.912	0.938	0.790
Tangibility (TG)	04	0.809	0.810	0.876	0.639

TABLE. IV. DISCRIMINANT VALIDITY - HTMT

	ATT	BI	EF	FN	IC	PEO	PU	DD	PR	SWOM	SN	TG
ATT												
BI	0.782											
EF	0.502	0.526										
FN	0.304	0.399	0.76									
IC	0.511	0.667	0.41	0.29								
PEOU	0.83	0.791	0.551	0.403	0.614							
PU	0.812	0.785	0.591	0.381	0.52	0.862						
DD	0.239	0.423	0.417	0.374	0.43	0.25	0.26					
PR	0.247	0.421	0.65	0.643	0.249	0.342	0.35	0.502				
SWOM	0.307	0.395	0.447	0.408	0.303	0.361	0.302	0.656	0.524			
SN	0.654	0.767	0.318	0.176	0.419	0.638	0.646	0.264	0.265	0.416		
TG	0.098	0.18	0.564	0.702	0.075	0.199	0.165	0.415	0.461	0.29	0.145	

TABLE. V. OUTER LOADING

	Items	Outer Loading		Items	Outer Loading		Items	Outer Loading
Attitude	AT1	0.893	Intrusiveness Concerns	IN1	0.879	Privacy	PR5	0.762
	AT2	0.92		IN2	0.929	Perceived Usefulness	PU1	0.893
	AT3	0.926		IN3	0.931		PU2	0.922
	AT4	0.935		IN4	0.891		PU3	0.939
Behavioral Intention	BI1	0.889	Perceived Ease of Use	PE1	0.905		PU4	0.957
	BI2	0.929		PE2	0.882		PU5	0.91
	BI3	0.947		PE3	0.91	Subjective Norms	SN1	0.848
	BI4	0.943		PE4	0.908		SN2	0.922
	BI5	0.965	Digital Dexterity	DD1	0.742		SN3	0.891
Efficiency	EF1	0.874		DD2	0.782		SN4	0.894
	EF2	0.882		DD3	0.861	Tangibility	TG1	0.717
	EF3	0.863		DD5	0.892		TG2	0.78
	EF4	0.876		DD6	0.862		TG3	0.852
Functionality	FN1	0.854	Privacy	PR1	0.888		TG4	0.841
	FN2	0.818		PR2	0.841	Social Electronic Word of Mouth	WM1	0.785
	FN3	0.757		PR3	0.87		WM2	0.806
	FN4	0.799		PR4	0.848		WM6	0.874

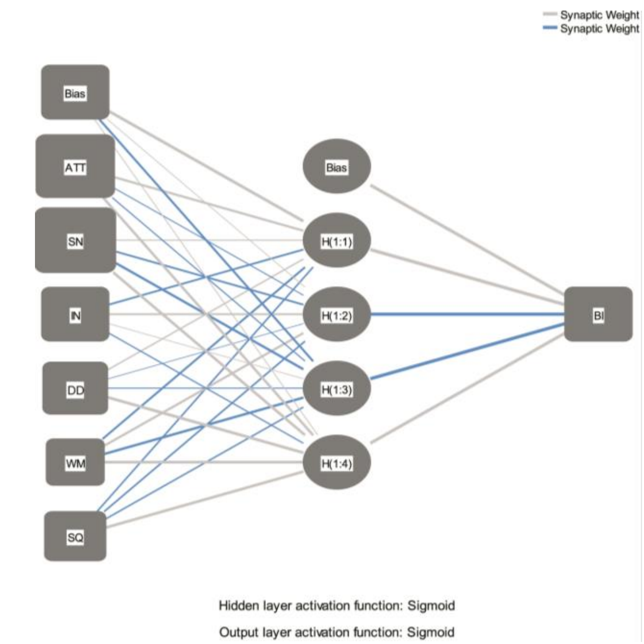


Fig. 4. ANN Diagram.

In SPSS 25, we conducted the Neural Networks test by multilayer perception tool. The exogenous variables, who had a direct impact by hypothesis on dependent variable (Behavioral Intention) were implied in covariate section with standardized rescaling. Partition of dataset was based on 70% training and 30% testing. The architecture was customized with one hidden layer. Sigmoid function was applied in both hidden layer and output layer with normalization correction 0.02. Batch training with optimization algorithm was applied as scaled conjugate gradient. Fig. 4 is depicting the synaptic weight of independent variables towards dependent variable. Normalized importance is the main tool for our study, as it shows the most important factor in terms of predicting the outcome of dependent variable. In our case, Subjective Norm (SN) is the best predictor followed by Attitude (ATT) with the highest rank. The least power of predictor is social electronic word of mouth (WM) while intrusiveness concern (IN), digital dexterity (DD) and service quality (SQ) have moderate prediction approach as presented in Fig. 5. The final output from neural network is prediction of scoring the behavioral intention on 5-point likert-scale towards adoption of smart mobility service. In Fig. 6, the chart showed the maximum points from agree to strongly agree scale.

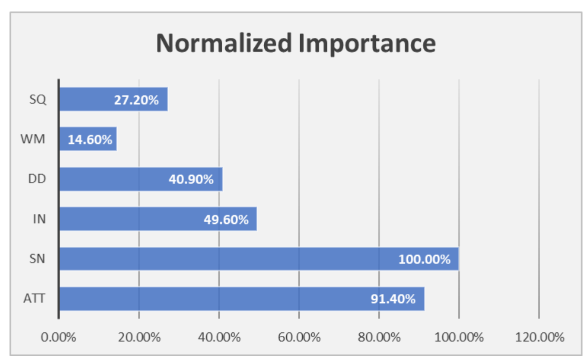


Fig. 5. redictors Normalized Importance.

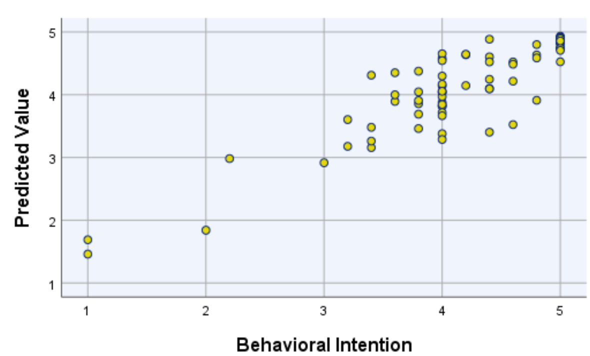


Fig. 6. Dependent Variable Prediction.

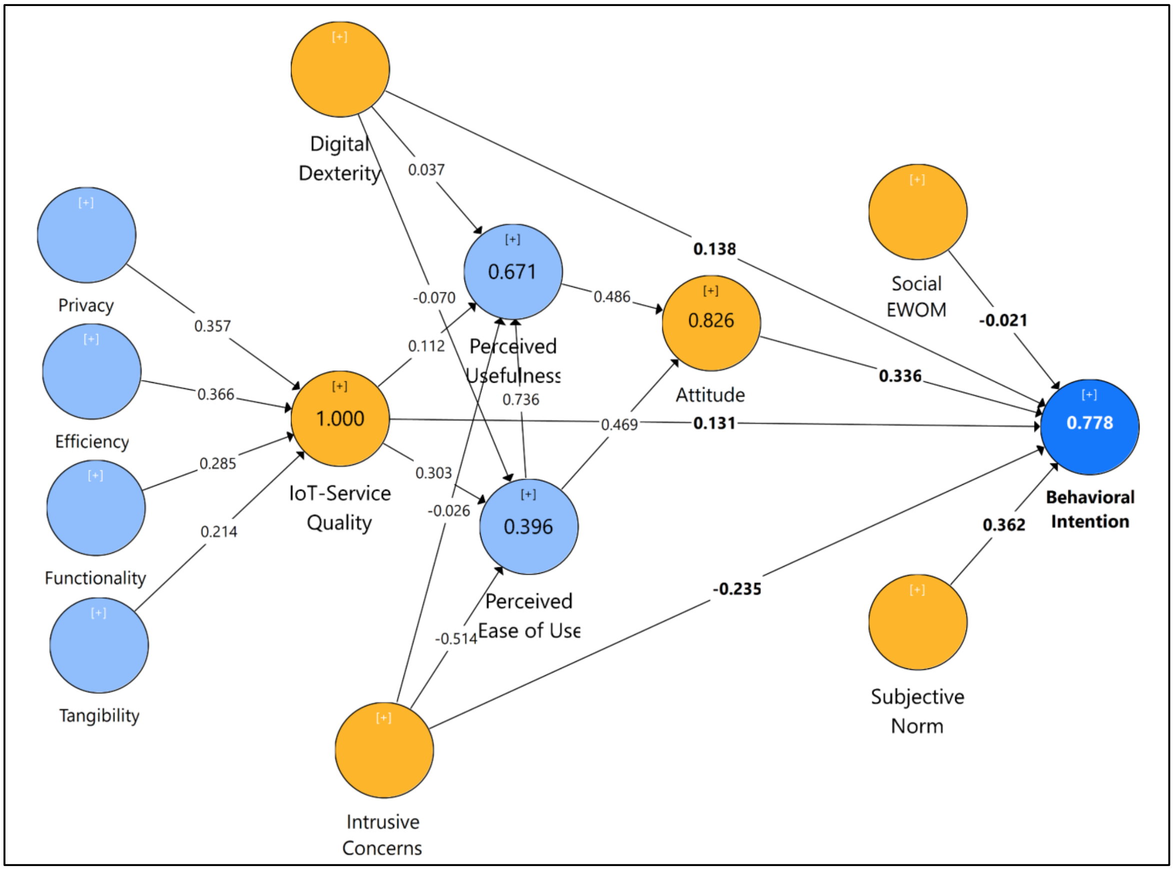


Fig. 7. Structural Equation Modeling (SEM) Causal Model.

SEM was also applied through SmartPLS software to realize the importance of constructs towards adoption of RFID service. SEM has multiple type of analysis i.e., Variance and Covariance based analysis. Although, variance-based analysis method used for pilot study data. The structural equation modeling of proposed model is illustrated in Fig. 7. The overall variance value is 0.778 that means factors in this model can predict 77.8% impact on behavioral intention towards the adoption of RFID service. The regression weights of all direct variables to BI are significant except SWOM. Subjective Norm emerged with the highest impact on BI (0.364). Attitude is appeared as significant influencer with 0.336 weight. Intrusiveness concerns showed the strong but negative impact on behavioral intention as hypothesized earlier. However, the least and negligible impact appeared from Social EWOM. The SEM analysis shows the model is well constructed with support of proposed hypothesis. As our all proposed hypothesis statements are proved true except one (i.e., Social EWOM and BI) in the preliminary results. This hypothesis will be considered to not include in final study or will identify the items to correct it [77].

V. Conclusion

The aforementioned analytical findings demonstrate the validation of our proposed research model. The results of instruments' reliability, validity, correlation, and outer loading evaluation proved the pilot survey authenticated for final study. The SEM-Neural approach has described the predictors of model that will support in finalizing the model for widespread survey. The relative importance in neural network is considered as major tool to take or drop the predictor in model. Merely relying on it does not guarantee that model is accurate or not. However, using SEM with Neural can explain the accuracy of model. As the least important predictor in Neural (i.e., WM) is appeared as insignificant in SEM. To ensure the precision and accuracy of the framework, SEM -Neural approach is recommended. In SEM approach, the impact of predictors on TAM variables (PU & PEOU) is considerable. The impact of main constructs (PU and PEOU) of TAM model on Attitude has shown the high variance of 0.826. The PU was hypothesized by 03 predictors (DD, IN, SQ) and variance calculated due to impact of these was 0.671 that is considered as higher. The empirical conclusion has resulted in reliable, valid and accurate theoretical framework for smart mobility service adoption in Malaysia.

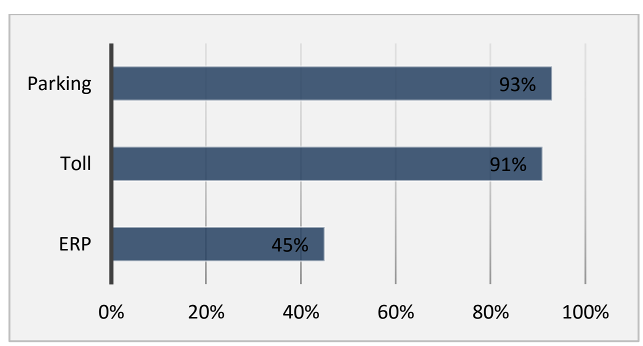


Fig. 8. RFID user Preference.

There is a huge potential for the growth of IoT based services in Malaysia, although such services are currently at initial stage. This pilot survey has concluded that higher number of drivers i.e., 93% are willing to use RFID for parking purpose. Use of RFID for tolling is preferred by 91% of respondents. While inclination to use RFID for Electronic Road Parking (ERP), only 45% respondent showed willingness as illustrated in Fig. 8. At this initial stage, the users' behavior to utilize RFID is portraying the image of future adoption pattern. The study will support authorities to devise the future policies regarding penetration of the RFID system. It will help in transforming the traditional transport system to Intelligent Transport System (ITS) with help of connected vehicles of RFID tag. It will also support in controlling congestion issue on urban roads. Furthermore, the Mobile-Wallet cashless payment method will integrate to make the things smart and intelligent. The given facts and figures of behavior preference towards smart mobility services will support the policymakers in service structure formation. It will also contribute to academic literature on smart mobility and Internet of Things in human computer interaction. It will assist the industry specific personnel for better understandings of user preferences. Lastly, the paper will contribute to the enduring research into finest technology adoption models that are pertinent for advancing the smart city contexts.

In this paper, TAM was extended with IoT and Digital environment perspective. The constructs IoT-Service Quality was introduced in this paper which had no traces in prior studies. To well-understand the human behavior in digital world, this paper introduced the new construct of Digital Dexterity in human-computer interaction literature. Practically, all pertained literature is cohered with the aspects that affect a users' decision to embrace IoT services with higher tendency of TAM. Furthermore, territory specific studies particularly in Malaysia, about the reasons that take the lead to the acceptance of Internet of Things services are almost scarce. A conceptual study was essential to comprehend the fundamental considerations and validate the analysis tools for full scale survey that encourage users to utilize IoT service. This paper emphasized the significance of various aspects on consumer intention to use IoT and to suggest it by propositioning a model by bringing into account technology adoption model and by expanding it with numerous new constructs. The proposed theoretical model is a provision to the existing literature since several uncharted variables have been included that will affect behavioral intent to make use of IoT services. This paper features the significance of pilot study regarding contribution to Information System literature and smart mobility context.

We propose to apply the validated instrument from this pilot study, conducted in Klang Valley, to the widespread survey with the large sample size across the main highways of Peninsular Malaysia. The results from the full-scale study will potentially contribute to the development of a context-specific model that can be employed to investigate smart mobility adoption and other technology diffusion in emerging economy. There is still unexplained variance in the model, that can be explained by adding new variables in the proposed framework. The framework has not encapsulated the system features in determining adoption behavior that could significantly impact the behavioral intention. As the research investigates RFID usage, the further studies could include the nature of RFID i.e., Passive RFID and Active RFID system in usability assessment. The hybrid analysis of SEM and Neural Network is the pioneer method in preliminary data analysis. However various empirical limitations of pilot survey such as covariance-based SEM analysis, detailed Neural Network processes and data normality factors can be overcome by large sample size. In future studies, various hybrid analysis techniques for predicting behavior in a pilot study such as System Dynamic Analysis (SDA), Association rule, fuzzification etc. could predict the more dynamic inferences.

References

[1] United Nations, "World Urbanization Prospects 2018 Highlights," 2018.

[2] A. S. Abdelfatah, M. Z. Shah, and O. C. Puan, "Evaluating the Sustainability of Traffic Growth in Malaysia," J. Traffic Logist. Eng., vol. 3, no. 1, pp. 6–11, 2015.

[3] Asia IoT Business Platform, "Mobility: A key aspect in Malaysia's Smart City Plans - Asia IoT Business Platform," 2019. [Online]. Available: https://iotbusiness-platform.com/blog/mobility-a-key-aspect-in-malaysias-smart-city-plans/. [Accessed: 30-Dec-2019].

[4] MAA, "Malaysian Automotive Association," 2017. [Online]. Available: http://www.maa.org.my. [Accessed: 25-Feb-2018].

[5] BCG, "Unlocking Cities: The impact of ridesharing in Southeast Asia and beyond 2017," 2017.

[6] D. Hong and L. W. Wong, "A study on smart mobility in Kuala Lumpur," Proc. 2017 2nd Int. Conf. Comput. Commun. Technol. ICCCT 2017, pp. 27–32, 2017.

[7] World Bank Group, "Malaysia Economic Monitor, June 2015 - Transforming Urban Transport," Malaysia Economic Monitor 2015, World Bank Group, 2015. [Online]. Available: http://www.worldbank.org/en/country/malaysia/publication/malaysia-economic-monitor-june-2015. [Accessed: 04-Mar-2019].

[8] M. Ali, S. Manogaran, K. M. Yusof, and M. R. M. Suhaili, "Analysing vehicular congestion scenario in Kuala Lumpur using open traffic," Indones. J. Electr. Eng. Comput. Sci., vol. 10, no. 3, pp. 875–882, 2018.

[9] J. Ali, T. Ali, Y. Alsaawy, A. S. Khalid, and S. Musa, "Blockchain-based Smart-IoT Trust Zone Measurement Architecture," in Proceedings of the International Conference on Omni-Layer Intelligent Systems - COINS '19, 2019, pp. 152–157.

[10] Nielsen Insights, "Cash Or Cashless? Malaysia's Shifting Payment Landscape," Nielsen Malaysia, 2019. [Online]. Available: https://www.nielsen.com/my/en/insights/reports/2019/cash-or-cashless-malaysias-shifting-payment-landscape.html. [Accessed: 07-Jan-2019].

[11] N. M. Noor, S. M. Sam, N. Firdaus, M. Azmi, R. Che, and M. Yusoff, "RFID-based Electronic Fare Toll Collection System for Multi-Lane Free Flow – A Case Study towards Malaysia Toll System Improvement," J. Telecommun. Electron. Comput. Eng., vol. 8, no. 4, pp. 71–76, 2016.

[12] S. P, V. Valsan, and P. C, "IoT Based RFID Gate Automation System," Int. J. Eng. Trends Technol., vol. 36, no. 9, pp. 471–473, 2016.

[13] S. Nandhini and P. Premkumar, "Automatic Toll Gate System Using Advanced RFID and GSM Technology," vol. 3, no. 5, pp. 13002–13007, 2014.

[14] J. Ali, A. S. Khalid, E. Yafi, S. Musa, and W. Ahmed, "Towards a Secure Behavior Modeling for IoT Networks Using Blockchain," in CEUR Workshop Proceedings, 2019, pp. 244–258.

[15] C.-L. Hsu and J. C.-C. Lin, "Exploring Factors Affecting the Adoption of Internet of Things Services," J. Comput. Inf. Syst., vol. 58, no. 1, pp. 49–57, 2018.

[16] L. Gao and X. Bai, "A unified perspective on the factors influencing consumer acceptance of internet of things technology," Asia Pacific J. Mark. Logist., vol. 26, no. 2, pp. 211–231, 2014.

[17] K.-L. A. Yau, S. L. Lau, H. N. Chua, M. H. Ling, V. Iranmanesh, and S. C. C. Kwan, "Greater Kuala Lumpur as a smart city: A case study on technology opportunities," 2016 8th Int. Conf. Knowl. Smart Technol., no. February, pp. 96–101, 2016.

[18] S. F. Verkijika and L. De Wet, "E-government adoption in sub-Saharan Africa," Electron. Commer. Res. Appl., vol. 30, no. May, pp. 83–93, 2018.

[19] J.-S. C. Lin and P.-L. Hsieh, "Assessing the Self-service Technology Encounters: Development and Validation of SSTQUAL Scale," J. Retail., vol. 87, no. 2, pp. 194–206, Jun. 2011.

[20] S. M. Hizam and W. Ahmed, "A conceptual paper on SERVQUAL-framework for assessing quality of Internet of Things (IoT) services," Int. J. Financ. Res., vol. 10, no. 5, pp. 387–397, 2019.

[21] Y. Feng and Q. Xie, "Privacy Concerns, Perceived Intrusiveness, and Privacy Controls: An Analysis of Virtual Try-on Apps," J. Interact. Advert., vol. 2019, pp. 1–41, 2018.

[22] S. K. Sharma, S. K. Mangla, S. Luthra, and Z. Al-Salti, "Mobile wallet inhibitors: Developing a comprehensive theory using an integrated model," J. Retail. Consum. Serv., vol. 45, no. July, pp. 52–63, Nov. 2018.

[23] Krishna Kumar, Sivashanmugam C, and A. Venkataraman, "Intention To Use Mobile Wallet : Extension of Tam," Int. J. Curr. Eng. Sci. Res., vol. 4, no. 12, pp. 5–12, 2017.

[24] Z. Mani and I. Chouk, "Drivers of consumers' resistance to smart products," J. Mark. Manag., vol. 33, no. 1–2, pp. 76–97, 2017.

[25] T. Natarajan, S. A. Balasubramanian, and D. L. Kasilingam, "Understanding the intention to use mobile shopping applications and its in fl uence on price sensitivity," J. Retail. Consum. Serv., vol. 37, no. January, pp. 8–22, 2017.

[26] R. Agarwal and J. Prasad, "A Conceptual and Operational Definition of Personal Innovativeness in the Domain of Information Technology," Inf. Syst. Res., vol. 9, no. 2, pp. 204–215, 1998.

[27] M. Iranmanesh, S. Zailani, S. Moeinzadeh, and D. Nikbin, "Effect of green innovation on job satisfaction of electronic and electrical manufacturers' employees through job intensity: personal innovativeness as moderator," Rev. Manag. Sci., vol. 11, no. 2, pp. 299–313, 2017.

[28] X. Zhou, Q. Song, Y. Li, H. Tan, and H. Zhou, "Examining the influence of online retailers' micro-blogs on consumers' purchase intention," Internet Res., vol. 27, no. 4, pp. 819–838., 2017.

[29] M. Barth, P. Jugert, and I. Fritsche, "Still underdetected - Social norms and collective efficacy predict the acceptance of electric vehicles in Germany," Transp. Res. Part F Traffic Psychol. Behav., vol. 37, pp. 64–77, 2016.

[30] N. Huete-Alcocer, "A Literature Review of Word of Mouth and Electronic Word of Mouth: Implications for Consumer Behavior," Front. Psychol., vol. 8, no. July, pp. 1–4, 2017.

[31] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," Manage. Sci., vol. 35, no. 8, pp. 982–1003, 1989.

[32] I. Ajzen, "The theory of planned behavior," Orgnizational Behav. Hum. Decis. Process., vol. 50, pp. 179–211, 1991.

[33] S. Pfoser, O. Schauer, and Y. Costa, "Acceptance of LNG as an alternative fuel: Determinants and policy implications," Energy Policy, vol. 120, no. April, pp. 259–267, 2018.

[34] B. Feng, Q. Ye, and B. J. Collins, "A dynamic model of electric vehicle adoption : The role of social commerce in new transportation," Inf. Manag., 2018.

[35] P. Taylor, E. Park, H. Kim, and J. Y. Ohm, "Understanding driver adoption of car navigation systems using the extended technology acceptance model," Behav. Inf. Technol., pp. 37–41, 2014.

[36] C.-K. Farn, Y.-W. Fan, and C.-D. Chen, "Predicting Electronic Toll Collection Service Adoption: An Integration of the Technology Acceptance Model and the Theory of Planned Behavior," Technol. E-government, 2007.

[37] H. Irawan, R. Hendayani, and D. Widyani, "Adoption of Electronic Toll Application Analysis," Int. J. Econ. Manag., vol. 10, no. S1, pp. 211–222, 2016.

[38] F. D. Davis, "Perceived Usefulness, Perceived Ease Of Use, And User Acceptance of Information Technology," MIS Quarterly; Sep, vol. 13, no. 3, pp. 319–240, 1989.

[39] A. A. Kervick, M. J. Hogan, D. O'Hora, and K. M. Sarma, "Testing a structural model of young driver willingness to uptake Smartphone Driver Support Systems," Accid. Anal. Prev., vol. 83, pp. 171–181, 2015.

[40] R. C. Jou, Y. C. Chiou, and J. C. Ke, "Impacts of impression changes on freeway driver intention to adopt electronic toll collection service," Transp. Res. Part C Emerg. Technol., vol. 19, no. 6, pp. 945–956, 2011.

[41] J. Globisch, E. Dütschke, and J. Schleich, "Acceptance of electric passenger cars in commercial fleets," Transp. Res. Part A Policy Pract., vol. 116, no. June, pp. 122–129, 2018.

[42] R. C. Jou, Y. C. Chiou, and J. C. Ke, "Impacts of impression changes on freeway driver intention to adopt electronic toll collection service," Transp. Res. Part C Emerg. Technol., 2011.

[43] M. J. Alsamydai, R. O. Yousif, and M. H. Al-Khasawneh, "The Factors Influencing Consumers ' Satisfaction and Continuity to Deal With E-Banking Services in Jordan," Glob. J. Manag. Businesss Res., vol. 12, no. 14, 2012.

[44] B. Alonso, R. Barreda, L. Olio, and A. Ibeas, "Modelling user perception of taxi service quality," Transp. Policy, vol. 63, pp. 157–164, 2018.

[45] A. Parasuraman, Z. Valarie A., and B. Leonard L., "Servqual : A Multiple-Item Scale For Measuring Consumer Perc," J. Retail., vol. 64, p. 12, 1988.

[46] A. Parasuraman, V. A. Zeithaml, and A. Malhotra, "E-S-QUAL a multiple-item scale for assessing electronic service quality," J. Serv. Res., vol. 7, no. 3, pp. 213–233, 2005.

[47] M. S. Janita and F. J. Miranda, "Quality in e-Government services: A proposal of dimensions from the perspective of public sector employees," Telemat. Informatics, vol. 35, no. 2, pp. 457–469, 2018.

[48] S. Hsu, F. Qing, C. Wang, and H. Hsieh, "Evaluation of service quality in facebook-based group-buying," Electron. Commer. Res. Appl., vol. 28, pp. 30–36, Mar. 2018.

[49] M. J. Alsamydai, "Adaptation of the Technology Acceptance Model (TAM) to the Use of Mobile Banking Services," Int. Rev. Manag. Bus. Res., no. 2014, pp. 2016–2028, 2016.

[50] E. Considine and K. Cormican, "Self-service Technology Adoption: An Analysis of Customer to Technology Interactions," in Procedia Computer Science, 2016, vol. 100, pp. 103–109.

[51] S. Boon-itt, "Managing self-service technology service quality to enhance e-satisfaction," Int. J. Res. Dev., vol. 7, no. 1, pp. 63–83, 2016.

[52] H. Boeck, J. Roy, F. Durif, and M. Grégoire, "The effect of perceived intrusion on consumers ' attitude towards using an RFID-based marketing program," vol. 5, pp. 841–848, 2011.

[53] Z. Mani and I. Chouk, "Drivers of consumers' resistance to smart products," J. Mark. Manag., vol. 33, no. 1–2, pp. 76–97, 2016.

[54] V. M. Wottrich, E. A. van Reijmersdal, and E. G. Smit, "The privacy trade-off for mobile app downloads: The roles of app value, intrusiveness, and privacy concerns," Decis. Support Syst., vol. 106, pp. 44–52, 2018.

[55] S. Hérault and B. Belvaux, "Privacy paradox et adoption de technologies intrusives Le cas de la géolocalisation mobile," Décisions Mark., no. 74, pp. 67–82, 2014.

[56] J. Lu, J. E. Yao, and C. S. Yu, "Personal innovativeness, social influences and adoption of wireless Internet services via mobile technology," J. Strateg. Inf. Syst., vol. 14, no. 3, pp. 245–268, 2005.

[57] R. Rothwell, Towards the Fifth ‐ generation Innovation Process, vol. 11, no. 1. 1994.

[58] K. ; Koivisto et al., "Extending the Technology Acceptance Model with Personal Innovativeness and Technology Readiness : A Comparison of Three Models," vol. 22, pp. 113–128, 2016.

[59] A. Turan, A. Ö. Tunç, and C. Zehir, "A Theoretical Model Proposal: Personal Innovativeness and User Involvement as Antecedents of Unified Theory of Acceptance and Use of Technology," Procedia - Soc. Behav. Sci., vol. 210, pp. 43–51, 2015.

[60] Z. Lin and R. Filieri, "Airline passengers' continuance intention towards online check-in services: The role of personal innovativeness and subjective knowledge," Transp. Res. Part E Logist. Transp. Rev., vol. 81, pp. 158–168, 2015.

[61] Chu Shu-chuan and Kim Yoojung, "Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites," Int. J. Advert., vol. 30, no. 1, pp. 47–75, 2011.

[62] N. B. Ellison and D. M. Boyd, Sociality Through Social Network Sites, vol. 1. Oxford University Press, 2013.

[63] M. Soderlund and S. Rosengren, "Receiving word-of-mouth from the service customer: An emotion-based effectiveness assessment'," J. Retail. Consum. Serv., vol. 14, no. 2, pp. 123–136, 2007.

[64] U. Gretzel and K. H. Yoo, "'Use and Impact of Online Travel Reviews,'" in Information and Communication Technologies in Tourism, 2008, pp. 35–46.

[65] A. Chong and E. Ngai, "What Influences Travellers' Adoption of a Location-based Social Media Service for Their Travel Planning?," in PACIS 2013 Proceedings, 2013.

[66] M. Ajzen, I. Fishbein, "Understanding Attitudes and Predicting Social Behavior. Prentice-Hall Inc., Englewood Cliffs, NJ.," Prentice-Hall Inc., Englewood Cliffs, NJ., 1980.

[67] M. N. K. Saunders, P. Lewis, and A. Thornhill, Research methods for business students, 7th ed. 2015.

[68] G. C. Moore and I. Benbasat, "Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation Stable URL : http://www.jstor.org/stable/23010883 Linked references are available on JSTOR for this article : of an Instrument to Measure the Percepti," Inf. Syst. Res., vol. 2, no. 3, pp. 192–222, 1991.

[69] R. E. Goldsmith and C. F. Hofacker, "Measuring consumer innovativeness," J. Acad. Mark. Sci., vol. 19, no. 3, pp. 209–221, Jun. 1991.

[70] S. Bambauer-Sachse and S. Mangold, "Brand equity dilution through negative online word-of-mouth communication," J. Retail. Consum. Serv., vol. 18, no. 1, pp. 38–45, 2011.

[71] I. Ajzen, "The Theory of Planned Behavior," 1991.

[72] World Population Review, "Kuala Lumpur Population 2018 (Demographics, Maps, Graphs)," World Population Review. 2018.

[73] M. N. K. Saunders, P. Lewis, and A. Thornhill, Research methods for business students. Pearson, 2012.

[74] 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," J. Acad. Mark. Sci., vol. 40, no. 3, pp. 414–433, 2012.

[75] Y. Owari and N. Miyatake, "Prediction of chronic lower back pain using the hierarchical neural network: Comparison with logistic regression—A pilot study," Med., vol. 55, no. 6, 2019.

[76] R. N. D'souza, P.-Y. Huang, and F.-C. Yeh, "Small Data Challenge: Structural Analysis and Optimization of Convolutional Neural Networks with a Small Sample Size," bioRxiv, p. 402610, 2018.

[77] J. F. Hair., W. C. Black, B. J. Babin, and R. E. Anderson, Multivariate Data Analysis William C . Black Seventh Edition. 2014.

Robotic Technology for Figural Creativity Enhancement: Case Study on Elementary School

Billy Hendrik[1], Nazlena Mohamad Ali[2*], Norshita Mat Nayan[3]
Dept. Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi, Malaysia[1, 2, 3]
Dept. Faculty of Computer Science, Universitas Putra Indonesia "YPTK" Padang, Padang, Indonesia[1]

***Abstract*—Robotic technology is a field that is in great demand today and it is very useful to human life, especially in the aspect of education. It would help students to be more active in the learning process. Creativity can be stimulating with the use of robotic technology. One kind of creativity is Figural Creativity (FC). This study investigated the effect of Robotic Technology as a learning tool to improve the FC skills of students. Forty (40) elementary school students aged 10-11 years, were the participants in this study. Students' creativity skills were measured from the Figural Creativity Test (TKF). This test was carried out before the intervention (pre-test) and after the intervention program (post-test). In the intervention program, students were given some education about robotic technology. To analyze the test results, we made use of the Statistical Service Products and Solutions package. The findings showed that the level of creativity in students with the K13 curriculum improved better, the FC scores of students in K-13 Curriculum were improved up to 23% with sig. 2-tailed = .000, p<.05 and the FC scores of the KTSP curriculum only improved only by 1.7 % with sig.2-tailed value = .572, p>.05. Thus, robotic technology learning is more effective in improving the FC of students with the K13 curriculum. Based on the result, we make a recommendation to the Ministry of Education that robotic technology is applied as an educational tool in the educational sectors.**

Keywords—Robotic technology; figural creativity; curriculum; TKF; education; KTSP; K-13

I. Introduction

Changes in the education sector have become a favorite topic for analysis in the Revolution Industry 4.0 era. Inevitably this change must be made to prepare the creative community to compete with other countries, because creativity is considered as a precious asset for solving problems and achieving sustainable development [1]. Changing in education systems would not be separated from changes of curriculum, learning methods and learning tools; In Indonesia, The Ministry of National Education has made changes to the curriculum several times, such as that of the KTSP (unit lesson-based curriculum) and K13 (character-based curriculum); these changes are aimed at developing the quality of education in this country [2]. Implementation of an education system based on curriculum K13 is expected to produce students who have creative and innovative skills [3]. However, in practice there may be some obstacles to the development of creativity, which might still not be optimal. According to Billy et.al (2018), there is no significant difference between the figural creativity level of students in KTSP curriculum with the figural creativity level of students in K13 curriculum; even the curriculum of KTSP influences student's figural creativity more [4]. In order for the objectives of the K13 curriculum to be achieved, it is necessary that there is improvement or development in its implementation.

To prepare children who will be ready to work in a world dominated by technology, we must develop the human creative spirit from the children as well as through technology. One of such technology that is growing today is robotics. A robot is a machine that can be programmed and reprogrammed to do certain tasks and usually consists of a manipulator such as a claw, hand, or tool attached to a mobile body or a stationary platform [5]. Robotic Technology on education will have a positive impact. Robotics offers a way to teach young children about the types of sensors and electronics they encounter in their daily life via a hands-on and engaging way. Teaching foundational programming concepts, along with robotics makes it possible to introduce children to important ideas that informs the design of many of the everyday objects they interact with [6].

Studies in the field of robotics have reported that robotics have a potential impact on students' learning in different subject areas (Physics, Mathematics, Engineering, Informatics and more) and on their personal development including cognitive, meta-cognitive and social skills, such as: research skills, creative thinking, decision making, problem solving, communication and team working skills, all of them being essential skills necessary in the workplace of the 21st century [7]. Educational robotics has emerged as a unique learning tool that can offer hands-on, fun activities in an attractive learning environment, feeding students' interest and curiosity. Robotic as a learning tool have the potential to positively contribute to successful learning [8]. Robot-based classes were found to increase creativity in an effective way, and all the sub-elements of creativity (fluency, originality, openness, and susceptibility) scored significantly higher in a post-verification test than the pre-verification test [9]. Creativity is the capacity to create, to produce new things, it is the capacity of the human brain to reach new conclusions and ideas as well as to solve problems in an original fashion [10]; another opinion "Creative thinking" is an original cognitive ability and problem solving process which enables individuals to use their intelligence in a way that is unique and directed toward coming up with a product [11]. Thus, Creative thinking skills is one of the capabilities required to solve various problems [12]. Creativity is increasingly recognized as a valuable asset for individuals in their daily problem solving and their professional careers, which contributes to personal and societal development [1], and being creative is a universal human attribute [13].

*Corresponding Author.

Furthermore, examined the use of a science and technology curriculum based on robotics to increase the achievement scores of youth ages 9-11 in an after-school program. The results revealed that youth in the robotics intervention had a significant increase in their mean scores on the post-test and that the control group had no significant change in scores from the pre-test to the post-test [7]. If we talk about innovation, of course, it will be related to creativity. Innovation is the implantation of creative inspiration. Innovation and creativity in the workplace have become increasingly important determinants of organizational performance, success, and longer-term survival [14], [15], [16]. However, the representation of creative abilities would have both local as well as global cultural influences [13]. Creativity is important to all businesses, whether they provide food, entertainment, transportation, or educational materials. It is essential to incorporate responsibility to the existence of any organization. Companies must be creative in order to compete in the marketplace and to continue developing products for a changing global society.

Although many studies have been carried out on the effect of robotics technology for creativity skills, nevertheless more studies still need to be conducted, especially with regards to the analysis of the application of robotic technology in different curricula and in the aspect of measuring the effects of robotics technology on creativity skills with standardized measuring instruments.

The goal of this study was to investigate the effect of Robotic Technology to improve the figural creativity skills of elementary school students who have different curricula. The results of this study may provide some vital suggestions to the educational sectors. This paper is divided into several sections. Section 2 discussed the materials and methods in carrying out the user experiment, followed by the results and discussion section.

II. Material and Methods

In this sub-chapter, we will explain the material and methods used in this study, such as participant, instrument, and procedure.

A. Participant

Participants in this study are elementary school students; with ages ranging from 10 to 11 years old, because development of creativity at this age is around 50%-70% [17], [18], [19], and at this age children are in the concrete operational stage, also in this stage the children has the ability to think rationally, imaginative, and can explore more objects or situation to solve problems. The participants were 40 students (N=40) from elementary schools with different curriculum in Sumatera Barat-Indonesia, KTSP curriculum and K13 curriculum. The students were selected by random sampling technique.

Total participants in the KTSP curriculum were 20 students (50%), and the total participants in the K-13 curriculum were 20 students (50%). The participant's mentioned was shortened to team A and team B as shown in Table I.

B. Instrument

Figural creativity of the students was measured by the Figural Creativity Test (TKF).Regarding the measurement of creativity, the most widely used test is the figural creativity test developed by Torrance known as TTCT (Torrance Tests of Creative Thinking) in the form of tests completing the picture (the drawing completions test) which is named the Wartegg test [20]. Other studies that also used TKF as a measure used it for the purpose of increasing the creativity of students through the implementations of 5E Learning Cycle with an Interrelationship Diagram for the students of X-8 class SMA Negeri 3 Surakarta school years 2012/2013 [21].

The creativity measured in TKF has an understanding as to the ability to form new combinations of given elements reflected in the fluency, flexibility, originality, and elaboration. Figural creativity scoring is comprised of four components: originality, fluency, flexibility, and elaboration [22], [23], [24]. The first element is originality, which implies that someone is able to produce an idea that is different from most people in the group. The answer is stated as original if the student is able to bring up a response that is included in 10% of the population, with a criteria such as the answer given by 10% or more of the subject is given a score of 0. If the answer is given by 5-9% of the subjects, then it would be given a score of 1. If the answer was given by 2-4% of subjects, then it would be assigned a score of 2. If the answer is given by less than 2% of the subjects, then it would be given a score 3. Answers that are not included in the list of originalities, were given a score of 3. The second element is fluency, which deals with measuring how a person is able to think and come up with ideas quickly, precisely, and varies with a large quantity because the number of responses determines the score (1 point per idea). The third element is flexibility, which is the ability a person has to be able to provide ideas from different points of view or activities, the score is the category idea (1 point per category). The fourth element is the elaboration, this is a way that a person has to develop, detailing and complete an idea or ideas (1, 2 or 3 points depending on the number of additional details drawn) [26]. The total score will be translated to a stan [25].

C. Procedure

In this study, in order to analyze the influence of robotics technology on creativity, we performed pre-test and post-test on elementary school students.

D. Pre-Test

Pre-Figural Creativity Test was performed on team A and team B in order to determine the level of figural creativity of the students before the intervention.

TABLE. I. Demography Table of Participant

Team	Curriculum	Total	Percentage (%)
A	KTSP	20	50 %
B	K-13	20	50 %

The pencil and paper tests were completed individually in a group administered session. The figural creativity test is made up of activities. The students are presented with a circular pattern (Fig. 1), and were asked to create an image which varies as much as possible and within 10 minutes it should have been determined using a circular pattern that has been given (Fig. 2); thus based on the images that have been made, the students were asked to provide the title of each image [26].

The total score is then being translated to the standard value and then the number of raw values becomes the creative quotient score, the creative quotient score also is being translated into four levels with intervals; Superior, High average, Average and Low Average [25]. The score for each creative quotient level can be seen in Table II.

E. Intervention of Robotic Technology

Robotic Technology intervention aims to increase the level of figural creativity in students, with the hope that the score of post-tests from students can increase. This section describes the activities that have been developed with children around the subject of robotics technology. The students from team A and team B got the intervention of robotic technology, this intervention consist of 7 meetings (see Table III).

Task 1. The robotics technology lessons consist of an introduction to the materials of the robotics technology via an audio-visual media, these activities give deeper knowledge about robotics technology. The students were introduced to several types of robots, such as robot with wheels, robot with legs, robot like a human. The next explanation was about the mechanical aspects of robots, sensor systems, microcontrollers, and actuators. Also, these explanations were being accompanied by experiment tools. The activity is shown in Fig. 3.

Fig. 1. Sample of Circular Pattern.

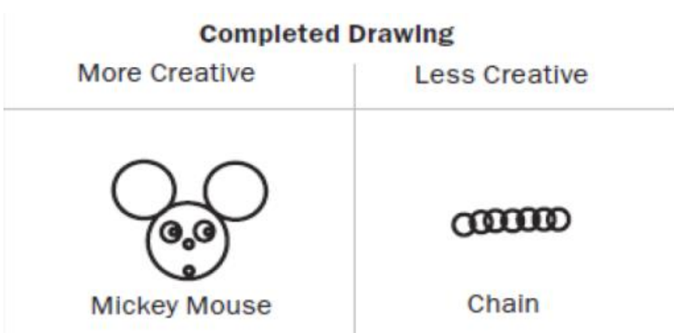


Fig. 2. Sample of Circular Pattern Answer Developed by Ellis Paul.

TABLE. II. LEVEL CREATIVE QUOTIENT

Level	Score
Superior	>= 120
High Average	111-119
Average	91-110
Low Average	80-90

TABLE. III. WEEKLY ROBOTICS TECHNOLOGY LESSON SCHEDULE

Lesson	Robotics Technology Lesson	Media	Purpose	Duration
Task 1 - First Lesson	Introduction and the materials of the robotics technology	Audio, video	Visualization and imagination give stimulation for fluency	2h. 15 min
Task 2 - Second Lesson	The robot development project using arm robot	Audio, video, robot, Module	Visualization, imagination, diverse classification give stimulation for fluency, flexibility (Module), elaboration (Module)	2h. 30 min
Task 3 - Third Lesson	The robot development project using mobile robot	Audio, video, robot, Module	Visualization, imagination, making stories, negation, modifying give stimulation for fluency, flexibility,	3h. 10 min
Task 4 - Fourth Lesson	The robot development project using drone	Audio, video, robot, Module	Visualization, imagination, negation give stimulation for fluency, flexibility (Module), originality (Module)	2h. 50 min
Task 5 - Fifth Lesson	The robot development project using Lego Mindstorms Track3r Tank Bot	Audio, video, robot, Module	Visualization, imagination, modifying, extending, concretizing, making stories, negation give stimulation for fluency, flexibility, elaboration, originality	3h. 20 min
Task 6 - Sixth Lesson	The robot development project using Lego Mindstorms Ev3rstorm Humanoid Bot	Audio, video, robot, Module	Visualization, imagination, modifying, extending, concretizing, making stories, negation give stimulation for fluency, flexibility, elaboration, originality	3h. 30 min
Task 7 - Seventh Lesson	The robot development project using Humanoid Robot.	Audio, video, robot, Module	Visualization, imagination, making stories give stimulation for fluency	3h. 30min

Task 2. After the introduction of the component, in the second meeting, the robot development project commenced. The students got new experiences with the robotic lesson using arm robot. Thus, the purpose of this activity is to provide knowledge to the students about static robot, one that

resembles the work of a human arm. Students will understand the motion of the robot system and also how to control the robot arm and its application. The activity is shown in Fig. 4.

Task 3. The students got some experiences from the mobile robot, this lesson was aimed at providing knowledge to the students about a robot that can move. Hence, the students can be able to distinguish between mechanical, actuator and control system in static and mobile robots. The activity is shown in Fig. 5.

Task 4. The students were introduced to the drone, this lesson aimed at providing knowledge to the students that the robot can fly. The Students can find out the mechanical, actuator and control system on the type of a flying robot. The activity is shown in Fig. 6.

Fig. 3. The Students were Introduced to Several Types of Robots, Such as Robot with Wheels, Robot with Legs, Robot Like a Human.

Fig. 4. The Students Get a New Experience with a Robotic Lesson using Arm Robot.

Fig. 5. Students Distinguish Mechanical, Actuator and Control System in Static Robots and Mobile Robot.

Fig. 6. The Students find out the Mechanical, Actuator and Control System on the Type of a Flying Robot.

Task 5. The students got some experience with Lego Mindstorms. Lego Mindstorms is an assembly kit that contains building block pieces (construction kits) and a programmable control unit that can enable one to build a number of robots [19], [20], [21]. This kit includes all the important components needed to build a robot, such as connectors, axle, busing, beams, frames, tubes, gears, belts, shafts, wheel, motors, sensors, and control center.

Control center in Lego Mindstorms is the Brick. The Brick can send a programme to the motors, receive information from sensors, among other functions. Additionally, the Lego Mindstorms consist of a large motor and a medium one as an actuator. It also comprises of different sensors: the color sensor, ultrasonic sensor, touch sensor, infrared sensor, gyro sensor, and the temperature sensor. This description was accompanied by displaying each of the components. The projects achieved by the students were, the making of a tracker tank bot, as the students were adequately taught on how to design, construct, build and control a tracker tank bot. The activity is shown in Fig. 7.

Task 6. The students also use Lego Mindstorms to make a Humanoid Bot. Students are able to know the differences between humanoid robots and other types of robots. The most important thing was that students can understand that Lego Mindstorms can form several types of robots, so students can develop their imagination to be able to create other types of robots. The activity is shown in Fig. 8.

Task 7. The students can also test platforms different from the original robot, the students were encouraged to use their imagination, and they were also introduced to control the robot with a smartphone, the robot used is the alpha robots.

Fig. 7. The Students Get Experience with Lego Mindstorms.

Fig. 8. Sample of Circular Pattern.

Fig. 9. The Students can Also Test Platforms different from the Original Robot.

After having conducting seven tasks, the process proceeded with the post-test in the control class. Also, the experimental class is expected from the intervention, this is so because 7 tasks in the classroom experiments can actually produce the post-test scores of the experimental class, which is much better than the control class. The activity is shown in Fig. 9.

F. Post-Test

Post-test is a measurement of the level of the figural creativity of students after getting intervention. After the intervention of the robotics technology was completed, the figural creativity of the students was measured by TKF (Post-test). Thus, Post-test needs to be done in order to determine the extent of the increase in the figural creativity of the student. The pencil and paper tests were completed individually in a group administered session. The figural creativity test in the post-test had the same activities as that of the pre-test. The total score were then translated to the standard value and the number of raw values becomes the creative quotient score, which was then translated into figural creativity levels.

III. Results

The case study was carried out on the elementary school with KTSP curriculum (Team A) and K-13 Curriculum (Team B). To analyze the results of the figural creativity test in pre-test and post-test, statistical analyses were conducted by using the Statistical Product and Service Solutions Software, and the data analysis is being divided into two parts.

A. Paired Sample T Test for Team A

The first part was the paired sample t-test, this analysis aimed to find out how significant the differences in the figural creativity of students were before the intervention and after the intervention. The first analysis is carried out on students with the KTSP curriculum.

H0: There are no differences in the mean values between the pre-test to the post-test of Team A, which means no intervention effects in improving the students' figural creativity.

Ha: There are differences in the mean values between the pre-test to the post-test of Team A, which means there are significant effects of intervention to improving the students' figural creativity.

Guidelines for decision making in paired sample t test analysis are based on significance values (sig) with the following provisions: if sig. (2-tailed) value < 0.05, it means that H0 is rejected and Ha is accepted, and then if sig. (2-tailed) value > 0.05, it means H0 is accepted and Ha is rejected.

Table IV shows the sig (2-tailed) value in team A = .572 P> .05; hence it can be concluded that there were no significant differences in the mean value of the figural creativity for the pre-test with the post-test, thus H0 is accepted and Ha is rejected. Paired differences mean = -1.7, these values show the difference between the mean value of pre-test (M = 99.35) with the mean value of the post-test (M = 101.05), the descriptive statistics is shown in Table V.

Based on the result of the figural creativity test in the pre-test and post-test for team A (KTSP curriculum), the student's scores were converted to figural creativity levels, thus the number of students found for each level of creativity in team A is shown in Table VI and Fig. 10.

TABLE. IV. Paired Samples Test for KTSP Elementary School

Team	Paired Differences		t	df	Sig. (2-tailed)
	Mean	Std. Deviation			
Pre-test team A - Post-test team A	-1.7	13.227	-0.575	19	0.572

TABLE. V. Descriptive Statistics in KTSP Elementary School Students

Team	Pre-Test				Post-Test			
	Min	Max	Mean	Std. Deviation	Min	Max	Mean	Std. Deviation
A	87	120	99.35	10.535	87	130	101.05	11.124

TABLE. VI. Pre-Test and Post-Test TKF Level in Team A (KTSP Curriculum)

TKF Level	Team A (Experimental Group)	
	Pre-Test TKF	Post- Test TKF
Superior	1	1
High Average	4	3
Average	9	10
Low Average	6	6
	20	20

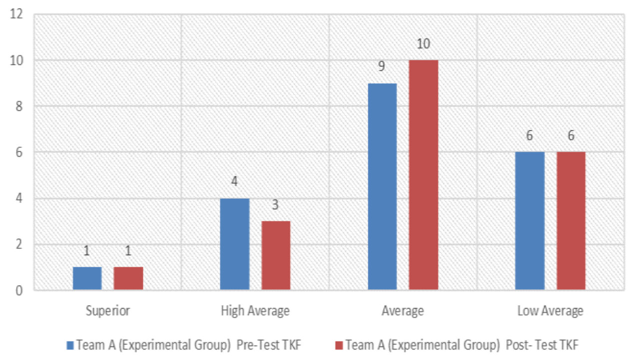


Fig. 10. Pre-Test and Post-Test TKF Level in Team A (KTSP Curriculum).

According to Table VI, the students in team A (N=20) consist of: superior level = 1 in pre-test and post-test, high average level = 4 students (pre-test) and 3 students (post-test), average level = 9 students (pre-test) and 10 students (post-test), low average level = 6 students in pre-test and post-test.

B. *Paired Sample T Test for Team B*

The second analysis was carried out on students in team B (K-13 curriculum). The following hypothesis is stipulated:

H0: There are no differences in the mean values between the pre-test to post-test of Team B, which means there are no intervention effects in improving students' figural creativity.

Ha: There are differences in the mean values between the pre-test to post-test of Team B, which means there are significant effects of intervention to improving students' figural creativity.

Table VII shows the sig (2-tailed) value in team B = .000 P< .05, so there were significantly different between the mean values of students' creativity before (pre-test) and after (post-test) the intervention, thus H0 is rejected and Ha is accepted. Paired differences mean = -21.85, these values show the difference between the mean value of pre-test (M = 116.80), the descriptive statistics is shown in Table VIII.

Based on the result of figural creativity test in the pre-test and post-test for team B (K-13 curriculum), the student's scores were converted to figural creativity levels, thus the number of students found for each level of creativity in team B is shown in Table IX and Fig. 11.

According to Table VIII, the students in team B (N=20) consist of: no student has figural creativity superior level in the pre-test but in the post-test, there were 9 students that had the figural creativity superior level; high average level = 1 student in the pre-test, and 6 students in the post-test; average level = 13 students in pre-test, and 5 students in the post-test; low average level = 6 students in pre-test and no student in the post-test; however, no student had low figural creativity level in both tests.

The next analysis was conducted on the mean value of students who received the intervention, this analysis aims to know the amount of the enhancement of figural creativity that is existent in each team (team A with team B)?

TABLE. VII. PAIRED SAMPLES TEST FOR KTSP ELEMENTARY SCHOOL

Team	**Paired Differences**		**t**	**df**	**Sig. (2-tailed)**
	Mean	**Std. Deviation**			
Pre-test team B-Post-test team B	-21.85	8.804	-11.1	19	0.000

TABLE. VIII. DESCRIPTIVE STATISTICS IN K-13 ELEMENTARY SCHOOL STUDENTS

Tea m	**Pre-Test**				**Post-Test**			
	Mi n	Ma x	Mea n	Std. Deviation	Mi n	Ma x	Mea n	Std. Deviation
B	85	11 3	94.9 5	7.584	10 0	13 3	116. 80	8.776

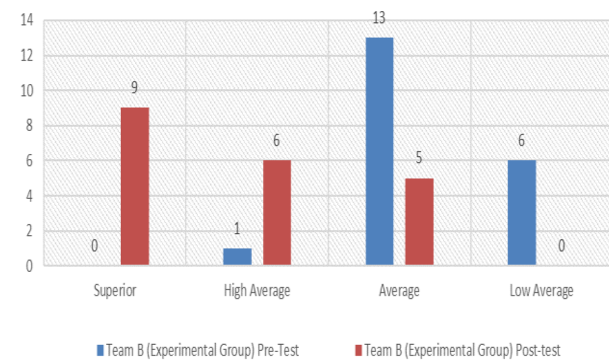


Fig. 11. Pre-Test and Post-Test TKF Level in Team B (K13 Curriculum).

The enhancement mean posttest of figural creativity ($\overline{X1}$) against the mean pretest ($\overline{X2}$) of figural creativity could be known by calculating the value of *P*. (*P* (%) = (($\overline{X1}$-$\overline{X2}$)/$\overline{X2}$) *x* 100%) [27].

$$P \text{ of Team A} = \frac{101.05 - 99.35}{99.35} x\ 100\%$$

$$= 1.7\ \%$$

$$P \text{ of Team B} = \frac{116.80 - 94.95}{94.95} x\ 100\%$$

$$= 23\ \%$$

The figural creativity of student's team A after getting the intervention of robotic technology were increased to 1.7%, while the figural creativity of student's team B after getting the intervention of robotic technology were increased up to 23%.

Based on Table X, in team A the changes in the number of students at each level of creativity in the pretest and posttest did not have a significant change, while in team B there were significant changes at each level, also in the pretest no student attained a creativity level superior level, but in the post-test there were 9 students; in the pretest only 1 student had a high average score, however in the post-test, there was an increase of up to6 students. Very interestingly, at the low average level, in the pretest there were 6 students that had low average of figural creativity, but in the post-test, the creativity decreased after the no-student intervention had a low figural creativity level.

TABLE. IX. PRE-TEST AND POST-TEST TKF LEVEL IN TEAM B (K13 CURRICULUM)

TKF Level	Team B (Experimental Group)	
	Pre-Test	**Post-test**
Superior	0	9
High Average	1	6
Average	13	5
Low Average	6	0
	20	20

TABLE. X. FIGURAL CREATIVITY LEVEL OF TEAM A AND TEAM B

TKF Level	Team A (KTSP-Experimental Group)		Team B (K13-Experimental Group)	
	Pre-Test TKF	**Post Test TKF**	**Pre-Test TKF**	**Post Test TKF**
Superior	1	1	0	9
High Average	4	3	1	6
Average	9	10	13	5
Low Average	6	6	6	0
	20	20	20	20

IV. DISCUSSION

Figural creativity is a very important skill possessed by students in the workplace of the 21st century, and the very interesting thing is that creativity skill can be developed by learning process, training, etc.[1].

The first finding of this study was by integrating robotic technology in learning activities; it has been proven to improve Figural creativity. The robotic technology learning is effective to improve figural creativity of students, this is consistent with several studies that have been done before, robotics is one way of fun learning which can make students think more creatively in creating something new [2]. Thus, robot can be used by children as a tool to boost new ideas and stimulate their creativity [3], also the potential of educational robotics has been acknowledged earlier, in particular with the potential to facilitate curiosity and creativity [4].

The second finding is that it is appropriate to apply robotic technology in the latest curriculum, namely the K-13 curriculum, based on the results of the figural creativity data analysis, which is much more improved for students with curriculum K-13; thus the use of robotic technology can support the purpose of K-13 curriculum learning process, which is to produce graduates who have creativity skills [5].

Therefore, we suggest that robotic technology as an educational tool can be applied in the educational sectors. This suggestion is consistent with De Zhang (2013) which strongly promotes robot innovative education, the educational reform, as well as promote the further development of robot technology [6]. Other suggestion is that, Robotic technology can be aligned with the students' curriculum, due to the fact that robotic project have showed a great upgrade in the children's education, in particular to develop general skills necessary in their life [7].

V. CONCLUSION

This study investigated the effect of applying technology in the enhancement of figural creativity between students with K13 curriculum and KTSP curriculum after intervening with them. To find out the initial level of the students' creativity, pretest was conducted for all participants. Researching the influence of technological intervention will provide suggestions regarding the development of student learning processes in schools to enhancement Figural Creativity skills. Based on the results of measuring the figural creativity of students from team A and B, there were unequal results from each team, which were from different experimental groups. This shows that the intervention of robotics technology in students with different curricula will result in creativity based on different improvements, especially the enhancement of the figural creativity skill of the students (see Table X).

Lesson one to seven is designed and arranged with good calculations, so students easily understand and are not burdened with something new. The lessons commenced with static robots to dynamic robots that can be controlled using a remote, even controlled using a smartphone. The material provided was arranged from the simplest to the higher technology. The robots used have various forms, colors, and functions that make students interested in learning and give them new knowledge and experience; with this, the students are able to discover or create new things in the field of technology in the future. This material follows the technological trends in the industrial revolution 4.0 era, so students will not be awkward with the technological advancements that exist. Lesson one to seven is packaged and delivered as seen in task 1 to task 7, intervention is done in 7 meetings. This material is delivered in audio, visual and kinesthetic forms, which is able to stimulate the four components of figural creativity, namely: originality, fluency, flexibility, and elaboration. The effect of this intervention is the figural creativity enhancement of the students.

Robotic technology can be an innovative educational tool that provides a positive effect on the development of the skill of students. Students who get the intervention (experimental group) have increased in their figural creativity better than students who did not get the intervention (control group). Based on statistical mean values of the figural creativity, the K13 curriculum in the experimental group achieved greater mean values of figural creativity than those in the KTSP curriculum, it can be concluded that robotic technology learning is more effective in enhancing the figural creativity in students with the K13 curriculum.

ACKNOWLEDGMENT

This research was supported by LLDIKTI from The Ministry of Research and Technology, which funded this research through the "Penelitian Strategis Nasional Institusi (PSNI)" program. We would also like to thank all participants for their time spent in this research.

REFERENCES

[1] E.Afari, & M. S. Khine, "Robotics as an Educational Tool: Impact of Lego Mindstorms", International Journal of Information and Education Technology, pp. 437–442, 2017.

[2] P. Alves-Oliveira, P. Arriaga, A. Paiva, and G. Hoffman, "YOLO, a Robot for Creativity," no. June, pp. 423–429, 2017.

[3] N. Anderson, & K. Potočnik, "Innovation and Creativity in Organizations : A State-of-the-Science Review," Prospective Commentary , and Guiding Framework. Journal of Management, vol. 40, no. 5, 1297–1333, 2014.

[4] J. E. Baek, "Effects of Robot-Based Learning on Learners ' Creativity," Education, vol. 127, pp. 130–134, 2016.

[5] B. Barbot, M. Besançon, & T. Lubart, "Creative potential in educational settings: its nature, measure, and nurture," Education 3-13, vol. 43, no. 4, pp. 371–381, 2015.

[6] F. Basoeki, F. D. Libera, E. Menegatti, & M. Moro, "Robots in education: New trends and challenges from the Japanese market," Themes in Science & Technology Education, vol. 6, no. 1, pp. 51–62, 2013.

[7] F. B. V. Benitti, "Exploring the educational potential of robotics in schools: A systematic review," Computers and Education, vol. 58, no.3, pp. 978–988, 2012.

[8] B. Hendrik, N. M. Ali, R. Sulaiman, M. Masril, and H. T. Fikri, "Intellectual Intelligence Analysis on Figural Creativity Levels for Elementary School Students," In Digital Transformation Landscape in the Fourth Industrial Revolution (4IR) Era pp. 167–176, 2018.

[9] D. T. Burhans and K. Dantu, "{ARTY:} Fueling Creativity through Art, Robotics and Technology for Youth," Aaai, pp. 4765–4770, 2017.

[10] H. Casakin, & S. Kreitler, "Evaluating Creativity in Design Problem Solving," Design Research Society, January 2014.

[11] T. Cassia, De, N. Pontifícia, U. Católica, & S. Paulo, "Intelligence and Creativity : Relationships and their Implications for Positive Psychology," Psico-USF, Bragança Paulista, vol. 20, no. 2, pp. 195–206, 2015.

[12] B. Cavaş, & T. Kesercioğlu, "The effects of robotics club on the students' performance on science process & scientific creativity skills and perceptions on robots, human and society," In 3rd International Workshop, Teaching Robotics, Teaching with Robotics, pp. 40–50, 2012.

[13] J. Chetty, "Lego©mindstorms: Merely a toy or a powerful pedagogical tool for learning computer programming?," Conferences in Research and Practice in Information Technology Series, vol. 159, pp. 111–118, January 2015.

[14] M. Fadly, "Increased CNC Learning Results," Universitas Negeri Semarang, 2015.

[15] M. Garaigordobil, "Intervention in Creativity With Children Aged 10 and 11 Years : Impact of a Play Program on Verbal and Graphic – Figural Creativity," vol.18, no. 3, pp. 329–345, 2006.

[16] M. Garaigordobil and L. Berrueco, "Effects of a Play Program on Creative Thinking of Preschool Children," Span. J. Psychol., vol. 14, no. 02, pp. 608–618, 2011.

[17] L. Hakim, "Analisis Perbedaan Antara Kurikulum," vol. 17, no. 2, pp. 280–292, 2017.

[18] B. Hendrik, N. M. Ali, R. Sulaiman, M. Masril, and H. T. Fikri, "The Effect Of Curriculum On The Level Of Figural Creativity Among Elementary School Students," Journal of Built Environment, Technology and Engineering, vol. 4, no. 11, pp. 200–205, 2018.

[19] S. Humble, P. Dixon, & E. Mpofu, "Factor structure of the Torrance Tests of Creative Thinking Figural Form A in Kiswahili speaking children: Multidimensionality and influences on creative behavior," Thinking Skills and Creativity, vol.27, pp. 33–44, 2018.

[20] R. Mangiwa, "Ability of Intelligence Quotient (IQ) Faculty of Medicine Students General Medicine Study Program," Jurnal E-Biomedik (EBM), vol. 2, no. 3, 2014.

[21] N. J. Marzieh Arefi, "Comparation of Creativity Dimensions (Fluency, Flexibility, Elaboration, Originality) between Bilingual Elementary Students (Azari language-Kurdish language) in Urmia City – Iran," The IAFOR International Conference on Language Learning, 2016.

[22] Sugito, Problems of Creativity And Measurement, Fakultas Bahasa dan Seni Universitas Negeri Medan, 1996.

[23] R., T. Herman, & J. A. Dahlan, "The Enhancement of Students' Critical Thinking Skills in Mathematics through The 5E Learning Cycle with Metacognitive Technique," IJER, vol. 4, no. 7, pp. 347–360, 2017.

[24] E. M. Nurhadi Santoso, "Pengembangan Kreativitas Siswa Berdasarkan Implementasi Kurikulum 2013," Pedagogik, vol. 2, pp. 19–25, 2014.

[25] A. A. Potur, & O. Barkul, "Gender and creative thinking in education: A theoretical and experimental overview," A|Z ITU Journal of Faculty of Architecture, vol. 6, no. 2, pp. 44–57, 2009.

[26] J. P. Rushton, "Performance on Raven ' s Matrices by African and White University Students in South Africa," vol. 28, no. 4, pp. 251–265, 2000.

[27] A. Vinsentricia, S. Sudarisman, & J. Ariyanto, "Efforts To Improve Students Figural Creativity Through the Use of 5E Learning Cycle With Interrelationship Diagram on Review," Pendidikan Biologi, vol. 7, no. 3, pp. 27-36, 2015.

The Impact of Deep Learning Techniques on SMS Spam Filtering

Wael Hassan Gomaa
Faculty of Computers and Artificial Intelligence
Beni-Suef University, Beni-Suef, Egypt, Shaqra University
Kingdom of Saudi Arabia

***Abstract*—Over the past decade, phone calls and bulk SMS have been fashionable. Although many advertisers assume that SMS has died, it is still alive. It is one of the simplest and most cost-effective marketing tools for companies to communicate on a personal level to their customers. The spread of SMS has led to the risk of spam. Most of the previous studies that attempted to detect spam were based on manually extracted features using classical machine learning classifiers. This paper explores the impact of applying various deep learning techniques on SMS spam filtering; by comparing the results of seven different deep neural network architectures and six classifiers for classical machine learning. Proposed methodologies are based on the automatic extraction of the required features. On a benchmark data set consisting of 5574 records, a fabulous accuracy of 99.26% has been resulted using Random Multimodel Deep Learning (RMDL) architecture.**

***Keywords*— SMS Spam Filtering; Deep Learning; RNN; GRU; LSTM; CNN; RCNN; RMDL**

I. Introduction

Telephone calls and mass text messages have been common over the past decade. For businesses to connect with their consumers on a personal level, it is one of the best and most cost-effective marketing tools. Internet and social media also reward the media, but SMS text messages are instant means that need few barriers to reach your audience with your information. Given that by 2020, seven out of every 10 people will own a smartphone, it will need to be easy to use and non-intimidating for any technology-based solution to achieve critical mass via digital. When companies want to keep up with this growing community, they need to continue looking for ways to make their company more mobile-friendly. In three seconds, more than 90% of people read an SMS, the chances of readability are extremely high. SMS may help increase customer engagement, promote products and services, or provide viewers with urgent updates. SMS messaging is also designed to enable companies to send messages to any contact number worldwide from any website or service using an API. Thanks to its omnipresent nature, SMS spamming has become a major nuisance for mobile subscribers. It includes substantial costs in terms of lost productivity, use of network bandwidth, administration, and personal privacy attack. Previous methods that detect SMS spam are hampered by the limitations of manual extraction, which is not efficient enough .Previous experiences and domain expertise are needed to identify prospective features for appropriate classification. Even then, it is important to reassess the selection of features on the basis of certain parameters, such as information acquirer. Deep learning is a category of artificial intelligence machine learning mechanisms. Deep neural network is a very effective way to avoid the wasteful phase of feature selection and extraction. For automatic pattern recognition and unsupervised feature learning, multiple layers of data are used. In order to automatically extract essential features and eliminate classification errors, the deep neural network components act with each other to train themselves sequentially. The effect of applying different deep learning techniques is discussed in this paper by comparing the accuracy of seven different deep neural network architectures and six classifiers for classical machine learning. The experiments are performed using Keras API and RapidMiner platform on a popular UCI benchmark dataset. The rest of the paper is structured as follows, Section II presents the related work, Section III describes the suggested methodologies, Section IV presents the experiments and findings and finally Section V outlines conclusion and future work stages.

II. Related Work

Several articles reviewed the previous and current approaches for SMS spam according to various metrics [1-7]. The most detailed and valuable review was presented in [1], it covers most of the SMS applications, approaches and methods. The authors in [1] searched in more than ten scientific databases such as Springer, IEEE, ACM and Google Scholar. They applied seven search strategies: approach-based, architecture-based, source-based, method-based, corpus-based, status-based and application-based. They found more than one thousand related articles; to exclude the obsolete and redundant papers, several sifting criteria were performed. For further research, eighty-three related articles were finally chosen. SMS spam detection methods are divided into three categories: machine learning, statistical analysis, and evolutionary algorithms. Naïve Bayes and decision tree are samples of machine learning algorithms. Statistical analysis techniques include models focused on mathematics, like factor analysis. Evolutionary algorithms are based on textual contents and biological techniques like genetic algorithms. The presented papers were categorized into three categories, content-oriented, non-content-oriented, and hybrid, according to approaches. The selected articles are classified into three categories according to architectures: client, server and collaborative. This systematic analysis presented many useful and interesting results for SMS spam filtering and concluded that much research remains to be done to improve existing approaches and methods [1].

There was another important review in [2]. The researchers in [2] presented a description of the problems, strategies and opportunities currently available on the role of SMS spam filtering. They introduced taxonomy of two main classes of methods and techniques: the access layer (AL) and the service provider layer. There are six sections in each class: Writing Style, Bayesian Network, SVM-based, Machine Learning, Evolutionary Algorithms, and Other Techniques. They compared 51 references in terms of datasets descriptions, proposed techniques, comparison techniques and major findings. They also concentrated on the weaknesses of existing studies and pointed out the paths for future research. Authors in [3] considered 17 papers and reviewed their algorithms, approaches, used databases, benefits, drawbacks, and methods of evaluation. In addition, they explained the classical machine learning classification problems. They concluded that the problems of shortcut terms and regional content were not addressed by any study. Authors in [4] concentrated on the various data mining strategies for the detection of SMS spam. Authors in [5] summed up recent efforts to reduce spam on SMS. In order to achieve better results, they recommended using the Support Vector Machine (SVM) classifier. Authors in [6] contrasted the algorithms, strengths, weakness, number of features, reliability and the data set used between 16 papers. Eventually, the first systematic review describing and comparing the major SMS spam identification processes, architectures and approaches was published in [7].

In addition, some research has been presented in [8-11] on deep learning approaches for the detection of SMS spam. Using text information only, CNN and LSTM were tested in [8]. On both balanced and imbalanced datasets, the experiments were performed. The results obtained indicated that the architecture of the CNN was superior to the architecture of the LSTM. Three models are presented in [9]: RNN, LSTM, and Semantic LSTM (SLSTM). SLTSM uses the semantic layer on top of the LSTM. The semantic layer is built using ConceptNet, WordNet, and Word2Vec.In [10] preprocessing of dataset was applied through stemming, sentiment analysis, stop word removing and tokenization. To be an input to CNN, a matrix of TF-IDF features was created. Authors in [11] tested CNN on two datasets containing respectively 5574 and 2000 SMS. The results of our suggested methodologies will be compared to the four researches mentioned above.

III. Proposed Methods

A. Classical Machine Learning Classifiers

In this section we discuss briefly the six used machine learning classifiers: Naïve Bayes, Generalized Linear Model (GLM), Fast Large Margin, Decision Tree, Random Forest, Gradient Boosted Trees and Support Vector Machine. A Naive Bayes classifier is a framework used for probabilistic machine learning classification tasks. It's easy to use but computationally cost effective. Naive Bayes' fundamental assumption is that, given the tag (class) value, the value of any attribute is independent of the value of any other attribute. GLM estimates models of regression for results after exponential distributions. These include the Poisson, binomial, and gamma distributions in addition to the Gaussian distribution. Each serves a different purpose and can be used either for prediction or classification depending on the choice of distribution and connection function. GLM is considered a dynamic version of linear regression models. Fast Large Margin is an SVM-Like algorithm which runs in O(N). Because of its complexity, Fast Large Margin is perfect for classifying big data. A decision tree is a flowchart-like architecture; it can be used to represent decision and decision-making visually and clearly. Each part of decision tree has a role in the classification process; the inner nodes represent checking of attributes, edges represent result of checking and the terminal nodes represent class labels. Random Forest model is developed from decision trees. The primary idea of a random forest is merging a number of decision-making trees into one model. Separately, the findings of decision trees may led to non-perfect results, but in combination, the findings are enhanced. Gradient Boosted Trees have the same idea of Random Forest models; but the difference is that in Gradient Boosted models the combination task starts at the beginning. If the parameters are carefully adjusted, gradient models may produce better results than random forests. The disadvantage of gradient models is that they suffer from noisy data. SVM is a popular and simple supervised classifier that depends on finding the hyper-plane which makes two given categories somewhat different. SVM is efficient in situations where the number of dimensions exceeds the number of instances.

B. Deep Learning Techniques

- Deep Neural Network (DNN): Older neural network models such as the first perceptrons were small, consisting of single input and single output layer, and at most single hidden layer between them. More than three layers count as "deep" learning (including input and output). It is a concept that is narrowly defined, meaning more than one hidden surface as shown in Fig. 1. Every layer of nodes trains in deep-learning networks on a different group of features according to the execution of preceding layers. The further you go into the neural net, the more complex the characteristics the nodes will identify as they integrate and recombine characteristics from previous layer.

- Recurrent Neural Network (RNN): RNN is a kind of artificial neural network with a "memory" that saves the previous information that is needed. The key aspect of RNN is Hidden State, which recalls certain knowledge about any given sequence such as word set in a sentence. The idea of RNN is to create many copies of the same architecture, each transmitting data to the next network, as shown in Fig. 2. Unlike other neural networks, it reduces the complexity of the parameters. RNN has many benefits, but suffers from the issue of vanishing gradient.

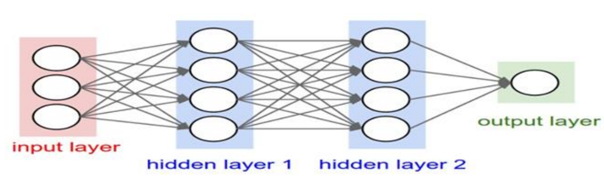


Fig. 1. Deep Neural Network.

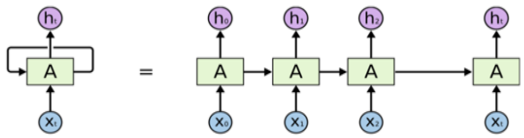


Fig. 2. Architecture of RNN.

- Long Short-Term Memory (LSTM): Several variants have been developed to resolve the problem of Gradients vanishing in RNN.LSTM is considered the best of them. In theory, a repeating LSTM system tries to "remember" all past information that the network has so far been seen and to "forget" irrelevant data. This is accomplished by adding different layers of activation functions called "gates" for different purposes. Each recurrent LSTM unit also preserves a vector called the Internal Cell State that determines the information that the previous recurrent LSTM unit has chosen to maintain conceptually. As shown in Fig. 3, LSTM contains four different gates: Input, Output, Input Modulation and Forget Gate.

- Gated Recurrent Unit (GRU): GRU is intended for solving the problem of the vanishing gradient that comes with a regular recurrent neural network. GRU is viewed as a variant of the LSTM. They have similar structures and deliver equally excellent results in some cases. This consists of only three gates, unlike LSTM, and does not retain an Internal Cell State. The data that is contained in an LSTM recurrent unit in the Inner Cell State is inserted into the Gated Recurrent Unit's secret state. The shared data will be passed to the next Gated Recurrent System. As shown in Fig. 4, Update Gate, Reset Gate, and Current Memory Gate are the different gates of a GRU.

- Convolutional Neural Network (CNN): Originally designed to conduct deep learning for computer vision tasks, the Convolutional Neural Network (CNN) has proven to be highly efficient. We employed the idea of a "convolution", a sliding window or "filter" that passes through the image, defining and evaluating important features one at a time, then reducing them to their essential features and repeating the process. Authors in [12] suggested CNN text classification techniques. Fig. 5 demonstrates a text-processing CNN architecture. It starts with an input sentence split into embedding words or words: low-dimensional representations created by models such as word2vec or GloVe. Words are divided into characteristics and fed into a convolutionary layer. The convolution results are either "pooled" or aggregated to a representative amount. This number is fed to a neural network that is completely connected, making a classification decision based on the weights assigned to each function within the text.

- Hierarchical Attention Network (HAN): HAN has been presented in [13] and is based on the same concept of the Attention GRU. HAN architecture is built using bi-directional GRU to acquire the word context. It contains two levels of attention for both word and sentence. Word2vec model is used to construct the required word embeddings. As shown in Fig. 6, HAN consists of several layers: word-encoder-layer, word-attention-layer, sentence-encoder-layer, sentence-attention-layer and fully-connected-layer.

- Recurrent Convolution Neural Network (RCNN): RCNN architecture is a combination of RNN and CNN to use the advantages of both techniques in a model. The authors in [14] proposed a text classification method based on RCNN as shown in Fig. 7. The main idea of this technique is capturing the required useful data by recurrent network and constructing the text representation using the convolutional structure.

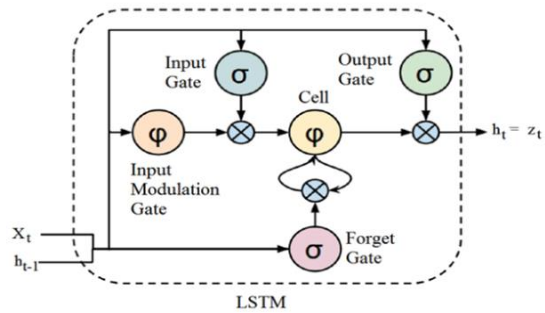


Fig. 3. LSTM Cell Architecture.

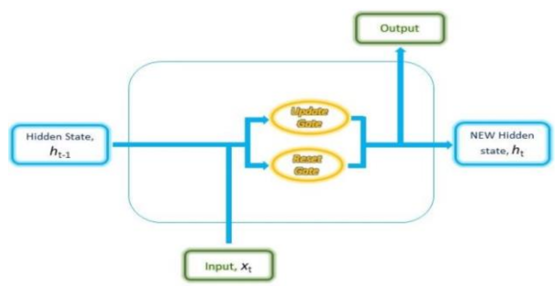


Fig. 4. GRU Cell Architecture.

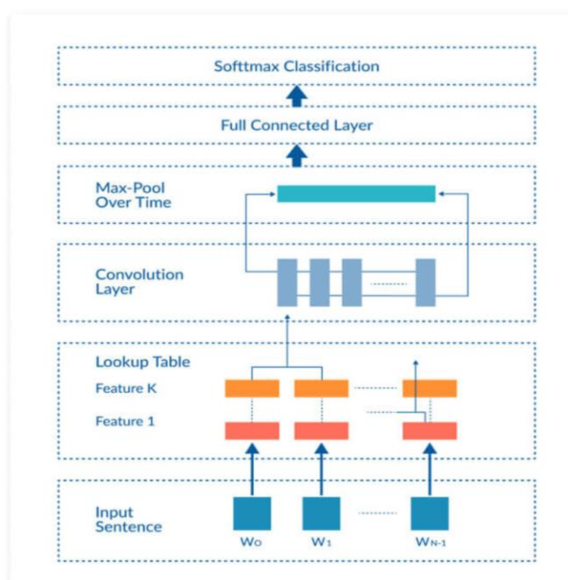


Fig. 5. CNN Architecture for Text Classification [12].

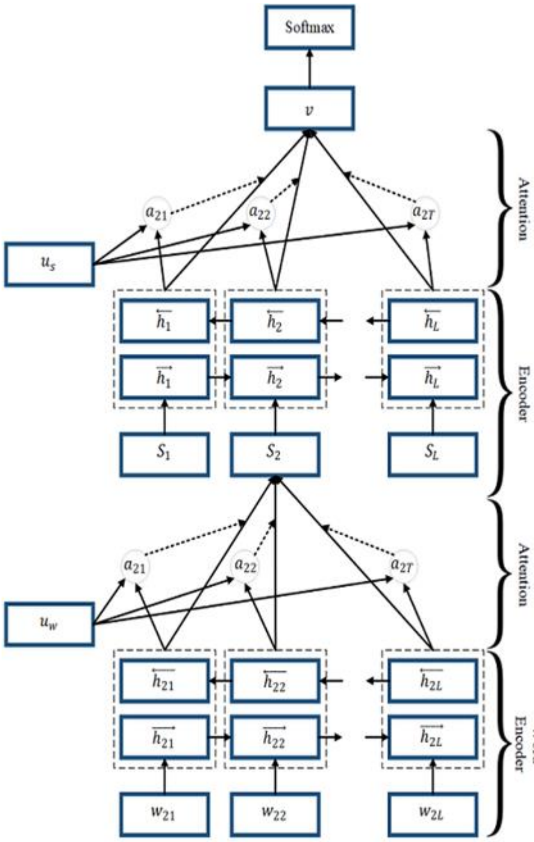


Fig. 6. HAN Architecture [13].

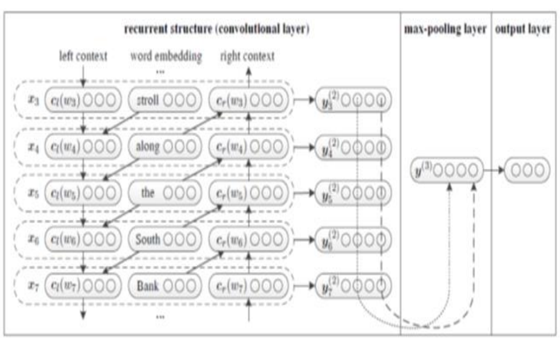


Fig. 7. RCNN Architecture for Text Classification [14].

- Random Multimodel Deep Learning (RMDL): Deep learning models across many areas have produced state-of-the-art tests. As shown in Fig. 8, architecture for classification of Random Multimodel Deep Learning (RDML) resolves the issue of getting the perfect design and structure from deep learning classifiers. RDMLs have the ability of accepting variety of input data including text, image, pictures and symbols. RMDL contains 3 random units, one left DNN classifier, one middle Deep CNN classifier, and one right Deep RNN classifier.

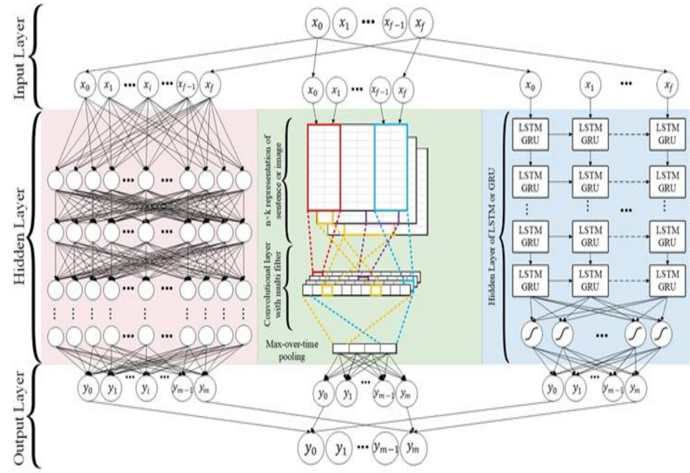


Fig. 8. RMDL Architecture [15].

IV. Experiments and Results

In this section, comprehensive results of the proposed methodologies will be discussed. The experiments were tested on the dataset available on UCI repositories for the collection of SMS spam. This dataset benchmark includes a total of 5574 English-language emails. Non-spam and spam numbers were respectively 4827 and 747[16]. Metrics like accuracy, precision, recall and F measure are used to assess the proposed methodologies. Table I shows the confusion matrix of Non-Spam and Spam SMS.

The various metrics are measured via the following equations:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{4}$$

All classical machine learning experiments described in Section III have been conducted using Auto Model; Auto Model is an expansion of RapidMiner Studio; it speeds up the model construction and validation process. This produces a mechanism that can be altered or created. This helps analyze the data given, provides appropriate models for problem solving, and helps to contrast the obtained results. All deep learning tests are performed using Keras and the python deep learning package. For all the deep learning experiments, the messages are converted into semantic word vectors using Glove model [17]. All networks begin with embedding layer that contains the semantic vectors resulted from Glove. Furthermore, Dropout layers are added to all deep learning models described in Section III to decrease the complexity of connections among the fully connected dense layers. The model of DNN contains 6 dense layers and 5 dropout layers. The model of LSTM contains 4 LSTM layers and 3 dropout layers in addition to the start embedding and final dense layers. The same structure is built for GRU model; it contains 4 GRU layers and 3 dropout layers. The architecture of CNN contains

7 convolution layers, 7 max pooling layers and 3 dropout layers. The architecture of HAN model has the same layers described in Section III [13]. RCNN model contains 4 convolution layers, 4 max pooling, 4 LSTM and 1 dropout layer. RDML model contains 3 DNNs, 3 CNNs, 3 RNNs as explained in Section III [15].

The findings of the proposed methods are presented in Table II and Fig. 9. Results showed that, relative to classical machine learning techniques, deep learning architectures have made significant improvements. Machine learning classifiers' highest accuracy is 96.86% according to Gradient Boosted trees. The highest accuracy resulted from RDML is 99.26 % of all possible methodologies.

TABLE. I. NON-SPAM AND SPAM CONFUSION MATRIX

	Non-Spam	**Spam**
Non-Spam	True-Positive (TP)	False-Positive (FP)
Spam	False-Negative (FN)	True-Negative (TN)

TABLE. II. THE RESULTS OF ALL PROPOSED METHODS

Proposed Classifier	**Accuracy**	**Precision**	**Recall**	**F Measure**
Naive Bayes	86.75	86.74	100.0	92.90
GLM	93.78	93.78	99.42	96.52
Fast Large Margin	95.10	95.04	99.49	97.21
Decision Tree	94.91	94.83	99.57	97.14
Random Forest	96.73	96.77	99.56	98.14
Gradient Boosted Trees	96.86	96.98	99.49	98.22
SVM	86.68	86.69	100.0	92.86
DNN	93.84	92.39	99.93	96.01
LSTM	97.88	97.82	99.56	98.68
GRU	98.12	98.23	99.44	98.83
CNN	98.24	98.25	99.57	98.90
HAN	99.17	99.09	99.65	99.37
RCNN	99.05	99.05	99.62	99.33
RMDL	**99.26**	99.25	99.59	99.42

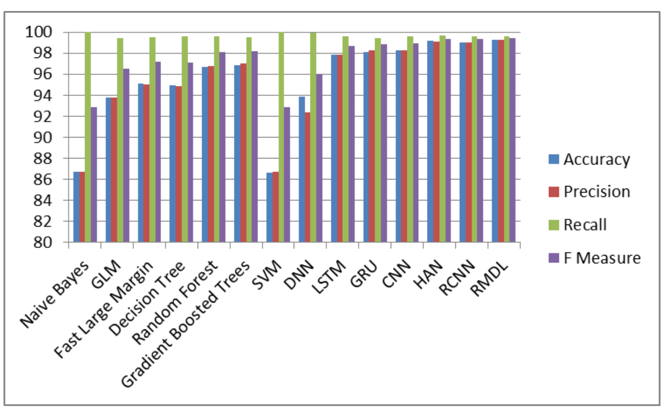


Fig. 9. Comparison of all the Proposed Methods.

TABLE. III. COMPARISON OF RMDL AND PREVIOUS DEEP LEARNING ARCHITECTURES

Classifier	**Accuracy %**
RDML	99.26
3CNN [8]	**99.44**
SLSTM [9]	99.01
CNN [10]	98.4
CNN [11]	99.1

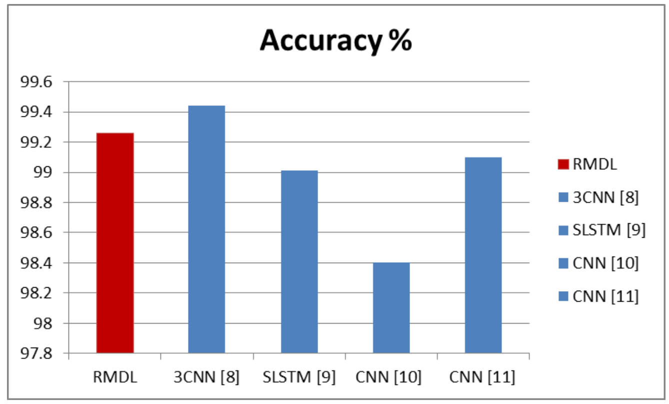


Fig. 10. Comparison of RDML and Previous DL Algorithms.

The main advantage of RMDL architecture is the power of finding the appropriate architecture for deep learning while at the same time reducing error rates and improving accuracy. The RMDL models include the advantages of the architecture of DNN, CNN and RNN. Table 3 and Fig. 10 provide a contrast between RMDL's accuracy and the best accuracy in the four deep learning articles [8-11] listed in the related work section. Using complicated 3CNN architecture in [8], the best accuracy was achieved.

V. CONCLUSION AND FUTURE WORK

This study covers SMS spam filtering task; thirteen proposed machine learning and deep learning classifiers were evaluated. Six classical machine learning classifiers were tested: Decision Tree, Naïve Bayes, Fast Large Margin, GLM, Random Forest, Gradient Boosted tress and SVM. Seven deep learning architectures were evaluated: DNN, LSTM, GRU, CNN, RCNN, HAN and RMDL. Benchmark dataset that contains 5574 messages was evaluated. The best accuracy of machine learning classifiers is 96.86 % resulted from Gradient Boosted trees. The best accuracy of all proposed methodologies is 99.26 resulted from RDML. These results indicate the power of using deep learning architectures on text classification modules. Our future work will focus on two main stages; the first stage is testing the model of Hierarchical Deep Learning for Text (HDLTex) [18]. The second stage is applying the techniques of transfer learning [19] that achieved promising accuracies on different classification tasks. of RMDL architecture is the power of finding the appropriate architecture for deep learning while at the same time reducing error rates and improving accuracy. The RMDL models include.

References

[1] Bayomi-Alli, O., Misra, S., Abayomi-Alli, A., &Odusami, M. (2019). A review of soft techniques for SMS spam classification: Methods, approaches and applications. Engineering Applications of Artificial Intelligence, 86, 197-212.

[2] Shafi'I, M. A., Latiff, M. S. A., Chiroma, H., Osho, O., Abdul-Salaam, G., Abubakar, A. I., &Herawan, T. (2017). A review on mobile SMS spam filtering techniques. IEEE Access, 5, 15650-15666.

[3] Lota, L.N., Hossain, B.M., 2017. A systematic literature review on SMS spam detection techniques. Int. J. Inf. Technol. Comput. Sci. 7, 42–50.

[4] Chaudhari, N., Jayvala, V., &Vinitashah, P. (2016). Survey on Spam SMS filtering using Data mining Techniques.International Journal of Advanced Research in Computer and Communication Engineering, 5(11).

[5] Kazi, K. F. I., &Dharmadhikari, S. C. (2014). Preserving the value of SMS texting: A survey on mobile SMS spam classification techniques and algorithms. Data Mining and Knowledge Engineering, 6(3), 106-112.

[6] Sajedi, H., Parast, G. Z., &Akbari, F. (2016). Sms spam filtering using machine learning techniques: A survey. Machine Learning Research, 1(1), 1-14.

[7] Delany, S. J., Buckley, M., & Greene, D. (2012). SMS spam filtering: methods and data. Expert Systems with Applications, 39(10), 9899-9908.

[8] Roy, P. K., Singh, J. P., & Banerjee, S. (2020). Deep learning to filter SMS Spam. Future Generation Computer Systems, 102, 524-533.

[9] Jain, G., Sharma, M., &Agarwal, B. (2019). Optimizing semantic LSTM for spam detection. International Journal of Information Technology, 11(2), 239-250.

[10] Popovac, M., Karanovic, M., Sladojevic, S., Arsenovic, M., &Anderla, A. (2018, November). Convolutional Neural Network Based SMS Spam Detection. In 2018 26th Telecommunications Forum (TELFOR) (pp. 1-4).IEEE.

[11] Gupta, M., Bakliwal, A., Agarwal, S., &Mehndiratta, P. (2018, August). A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers. In 2018 Eleventh International Conference on Contemporary Computing (IC3) (pp. 1-7). IEEE.

[12] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[13] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., &Hovy, E. (2016, June). Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 1480-1489).

[14] Lai, S., Xu, L., Liu, K., & Zhao, J. (2015, February). Recurrent convolutional neural networks for text classification.In Twenty-ninth AAAI conference on artificial intelligence.

[15] Kowsari, K., Heidarysafa, M., Brown, D. E., Meimandi, K. J., & Barnes, L. E. (2018, April). Rmdl: Random multimodel deep learning for classification. In Proceedings of the 2nd International Conference on Information System and Data Mining (pp. 19-28).ACM.

[16] Almeida, T. A., Hidalgo, J. M. G., &Yamakami, A. (2011, September). Contributions to the study of SMS spam filtering: new collection and results. In Proceedings of the 11th ACM symposium on Document engineering (pp. 259-262).ACM.

[17] Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[18] Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S., & Barnes, L. E. (2017, December). Hdltex: Hierarchical deep learning for text classification. In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 364-371).IEEE.

[19] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification.arXiv preprint arXiv:1801.06146.

Translator System for Peruvian Sign Language Texts through 3D Virtual Assistant

Gleny Paola Gamarra Ramos[1], María Elisabeth Farfán Choquehuanca[2]
Ingeniería de Sistemas
Universidad Nacional de San Agustín de Arequipa
Arequipa, Peru

***Abstract*—The population with hearing impairment in Peru is a community that doesn't receive the necessary support from the government, nor does it have the necessary resources for the inclusion of its members as active persons in this society. Few initiatives have been launched to achieve this goal and for this reason we will create a resource that will give the deaf community to have greater access to textual information of the listener-speaking community. Our goal is to build a tool that translates general texts and also academic content such as encyclopedias into Peruvian sign language which will be represented by a 3D avatar. This translation will feature different lexical-syntactic processing modules as well as the disambiguation of terms using a small lexicon, similar to Wordnet synsets. The project is developed in collaboration with the deaf community.**

***Keywords*—Peruvian sign language; Lexical-Syntactic Analysis; Avatar 3D**

I. Introduction

The ability and need to communicate with each other is a fundamental part of the human being. There are almost 7,000 different languages around the world. As our world connects more and more as technology advances, language translation provides a critical cultural bridge between people from different countries and ethnic groups. This is the case of the deaf community.

The sign language community is of considerable size around the world, which is why a lot of tools are made to achieve the inclusion of these people within society, our country is no stranger to this reality, as evidenced by the surveys carried out by the INEI (Instituto Nacional de Estadıstica e Informática) [1].

According to surveys carried out by the INEI [1], it has been show that there are more than half a million deaf people in Peru, it is an average percentage, considering that in the rural parts of the country many people suffer from this disability however they don't belong to any recognized deaf society or community, so they have no knowledge of the Peruvian Sign Language (LSP) [2], due to the few resources they have, and the lack of dissemination of the community. Members of the deaf community that make use of the LSP is make up a smaller group.

The Peruvian Government published a Guide for learning LSP, a book to support the disabled, seeking to strengthen sign language. But this is by far not enough because all educational texts of regular schools are written in the Spanish language and its translation causes students with this disability to have difficulties acquiring knowledge because the texts aren't interpreted in their language. The great problem that exists when translating a text into sign language is the semantic plurality they present, which differ from place to place, each country has its own LS (Sign Language) [3], [4] and in Peru it is called, Peruvian Sign Language-Lengua de Señas Peruano (LSP).

In our country, attention to the LSP community has been neglected by the authorities, despite the fact that in the political constitution of Peru, article 7 [5], declares that it is the duty of the Peruvian government to protect this community, but however this is not being fulfilled according to the results established by the congress of the republic.

The problem faced by deaf students in Peru every day is that they don't have their own educational material in their language, making this their main obstacle to improving their educational quality, this because the regional national curriculum of Arequipa [6], [7] doesn't contemplate them within it explicitly, thus having a great difference in educational quality compared to other students. Currently, only 6 deaf students attend the National University of San Agustín de Arequipa (UNSA) [8], the first public institution to provide support to these students, providing them with an accompanying staff throughout the course of their university preparation, to translate all the lessons.

Our objective in carrying out this work is to collaborate significantly with the deaf community of Peru, because they don't have their own language educational material, that allows them to improve their learning process. This system will allow them to translate academic texts by improving their educational quality at various levels: Both in primary, secondary and higher education.

This work is organized as follows: In Section II we will find the various related investigations, section III and IV concepts price and methodology, section V analysis of results, section VI and VII conclusions, future work and acknowledgments.

II. Work Related

The first steps in Natural language processing (PLN), which were taken in the 1960s, were basically the automatic translation of texts, but there were always certain conflicts about linguistic contexts and semantics. Thus, over the years, researchers focused more on the syntactic analysis, so the main objective was to improve the ambiguities about the context, which was initially presented.

With the revolution of the artificial intelligence in the last years the processing of the natural language is more and more important, this generates that systems of different areas are developed [9]. Meanwhile, sign language has been around the world since ancient times. In Peru [5] in 1987 the government published and recognised Decree Law 29535 standardising it with the Manual LS and reprinted it in 1996, on a par with several other countries, to enable the creation of applications to collaborate with this community.

In 2015, in Peru, a system was developed for the first time that collaborated with the inclusion of the people of this community, so it seeks to improve their quality of life, allowing them to have better communication and interaction with society [10].

In 2016, SAILCA (Sistema de Interpretación Automática del Lenguaje de Señas para el Cálculo de la Enseñanza) was developed in Colombia [11], [12] becoming one of the most influential tools due to the importance of the subject, in addition to developing its own language, thus promoting respect for the diversity of hearing impaired communities.

Peluso [3] mentions the development in TUILSU (University Technician of Interpretation LSU-Spanish), University of the Republic, on the development of a project for the elaboration of an LSU dictionary (Uruguayan Sign Language).This dictionary relates the filming of each sign to the filming of its respective meaning.

In Costa Rica [11] a project was carried out to formalize the Costa Rican Sign Language (LESCO) of the National Resource Center for Inclusive Education (CENAREC), in order to create a text translator to LESCO using JSON. The prototype presented for this project made use of 3D avatars and was developed for a web environment.

Tracey in [17] builds a deep neural network that serves as a tool that receives English text as input and generates a text in French as an answer. Reaching the highest possible translation accuracy.

Raziq in [18] seeks to solve the problem that is presented in Pakistan with deaf-mute children, in this way they seek to include them in society, for this purpose they propose a sign language translator from Pakistan where they use Leap technology motion. This device allows them to obtain greater precision in the recognition of the movements made by the hands and the figures they are forming. They seek to train each of these movements in sign language so that they are subsequently entered and analyzed by a correlation algorithm, in order to recognize the movement and thus be able to convert it into text.

III. Peruvian Sign Language

Peruvian Sign Language is based on the French Sign Language (LSF) and has a grammar following the model: Subject - Verb - Object [14]. Next, we will see three important elements of the LSP.

A. Grammar

The LSP maintains a Subject - Verb - Object (SVO) structure. For example, the translation for the sentence:

"Yo voy a mi casa"

Translating word for word would be obtained as follows:

[YO] [IR] [CASA]

B. Articulators

They are tools that a language uses to express the words in an understandable way. In the LSP, the hand is mainly used to sign the same part of the body from the waist up. According to [15] there are 25 toponemas that are the regions of the body that are involved in the formation of signals.These articulators can be seen in Fig. 1.

C. Word Formation

There is in the sign language a way to build words letter by letter which is called typing, which is useful when writing proper names or words whose signal is unknown. The signs are coined to a word or phrase due to common use and any of the actuators (one or two hands, face, hand movement) can intervene in its formation.

-If we say "Sali", the translation will be:

[SALIR] [PASADO].

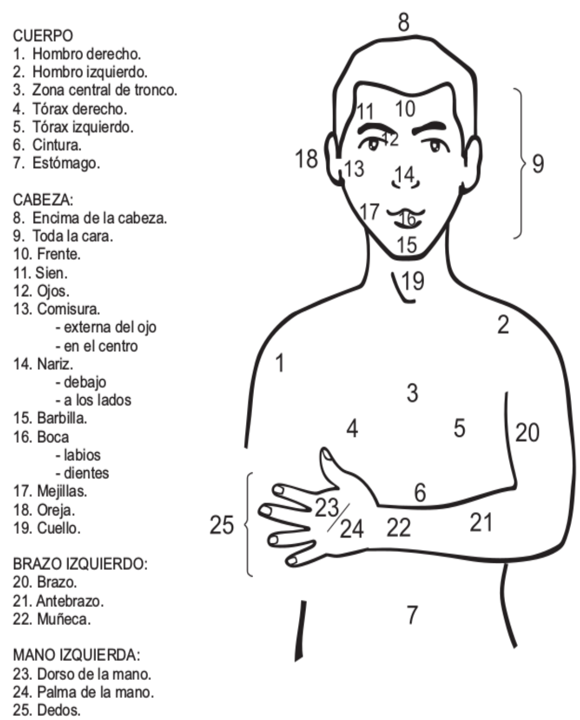


Fig. 1. Toponemes, Source[12].

IV. Virtual Assistant

For the virtual assistant, the capture of movements of a translator in LSP was first considered, which will be emulated by a 3D avatar. This movement capture is done with support tools, first the use of Kinect was considered due to previous investigations.

In [13], [14] the use of Kinect to capture movements of a person and to extract them is explained, previously this with a support software, In the analysis of the planned architecture, Iclone was defined as said software that provides support for a better management of the final capture obtained.

Iclone allows connection to Kinect, that allows the movement to be captured and encapsulated in a chosen video format, but still performing the motion capture. For the signs it is necessary that the capture of the movements of the gestures of the face and fingers are well defined, due to the expressive language that some signs require. Therefore, the use of Leap Motion was included, thus allowing a better development of the avatar.

a) IClone: iClone is a software of rendering and animation in 3D[15], allowing users to make animations as a movie. This software allows the inclusion of external tools for better operation of capture and movement representation. Fig. 2 shows the Iclone platform in the process of analyzing the structure of the avatar to be used.

b) Kinect: This device was developed for the Xbox 360 console and shortly after being on sale and after the publication of several versions of unofficial drivers, Microsoft decided to release the development API[14].Figure 3 shows the structure of kinect.

The use of Kinect [13] for an external capture of a sign translator is important in this project because it allows the first main capture of the translator to emulate. Kinect from Microsoft who after a few years decided to release the development suite of the SDK (Software Development Kit) device has versions that have improved its compatibility with various devices, such as Leap Motion.

For motion capture and avatar modeling, Kinect allows the creation of the avatar based on the articulated structure supported by it, so that, from the avatar mesh, fits all your skeleton, focusing on the joints that are necessary for the capture system, conditioning to the needs of the signs.

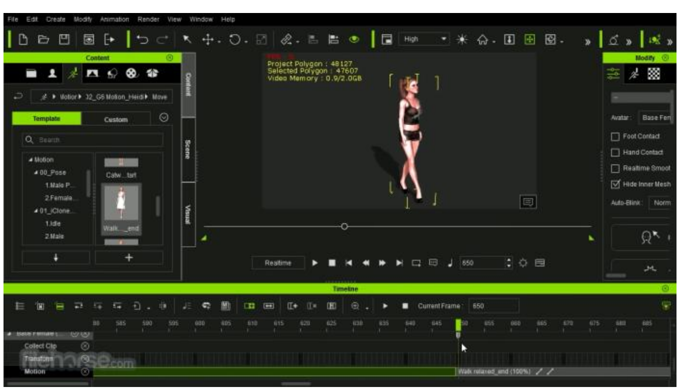


Fig. 2. IClone Platform. Source [15].

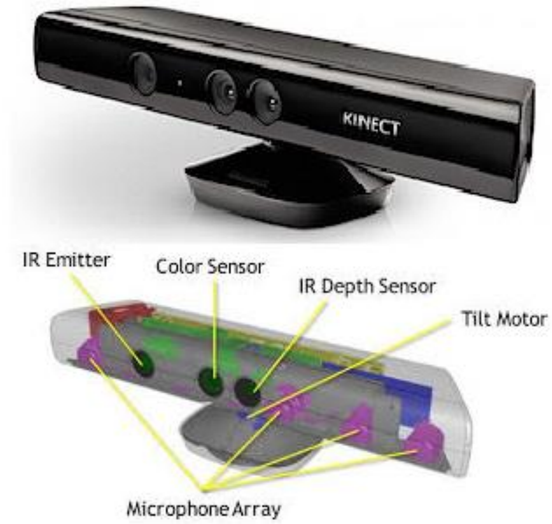


Fig. 3. Kinect. Source[14].

c) Leap Motion: In [16] Leap Motion is described as a small device that allows you to control your computer based on gestures that are generated with the movement of your fingers. One can create a virtual image trace of the hands and joints from the wrist by tracking all movements.

In Fig. 4 and Fig. 5 we can see the leap motion and the capture of the structure of the skeleton of movements that it recognizes in hands. This allows us to improve in details of hands and be more precise in the translation.

Fig. 4. Leap Motion Device Source[16].

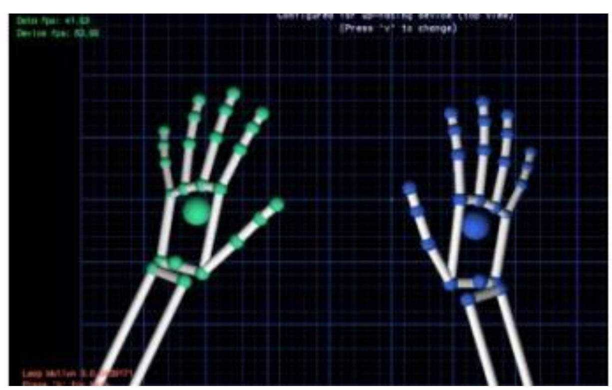


Fig. 5. SDK Leap Motion Display Source [16].

V. Translation of Texts

a) Natural Language Processing (NLP): Natural Language Processing according to [19], [20]. It is the discipline that arises from the communication between man and machine. It is more fluid, so that in this way the machine adapts to human language and not vice versa, also the information expressed in a human language is used through computer systems. Natural language processing is polysemantic, generating certain difficulties in the formalization of grammar.

There are various types of NLP [20], [21] What we consider are the following: The lexicon, morphological, syntactic and semantic.

b) Lexico Analysis: In [22] it is indicated that the lexical analyzer specifies the tokens of the language considering that there may be several tokens that correspond to the same regular expression.

c) Morpho Syntactic Analysis: Also called parsing, wherein [23] indicates that it is the combination of the recognition of an input chain with the assignment to it of a syntactic structure. It provides the tools to form words from smaller units.

d) Syntactic Analysis: Establish how the combination of words should go to form valid sentences and study how they relate to each other. Performs the analysis of the syntactic structure of the phrase through a grammar of the analyzed language [24], [25]. The figure below shows the shallow parsers that are responsible for identifying the structures as a noun phrase, Verbal and prepositional.

VI. Materials and Methods

The proposed architecture that was established to carry out this project is as follows:

In Fig. 6 the architecture established for the system is observed, which allows us to observe the process that will be carried out to achieve the translation and that it is represented by a 3D avatar.

The StanfordNLP was established to use natural language processing [26], which provides us with a set of human language technology tools. We make use of this tool for the named entities recognition (NER), established in the corpus.

The use of WordNet (WN) [27], which is a lexical database with a wide variety of languages, was also determined. It was created by Princeton University and represents a conceptual and structured semantic network. WordNet defines names, verbs, adjectives and adverbs. The basic unit of information in WN is the synset (synonym sets or synonyms sets).

As can be seen in Fig. 6 and Fig. 7, the corpus is obtained and subsequently the corresponding labeling of the named entities (NER) is carried out, with the help of the Stanford tool. A grammar model is made, and each sentence is analyzed lexically and syntactically, resulting in sentences with morphosyntactic words.

Later, an analysis is made to identify the synsets, for those sentences that have nouns, adjectives and verbs. Applying the python nltk library, for the proper handling of ambiguities, which happens with words that have the same meaning. We make the connection with the wordNet database, identifying each of the words achieving the corresponding translation.

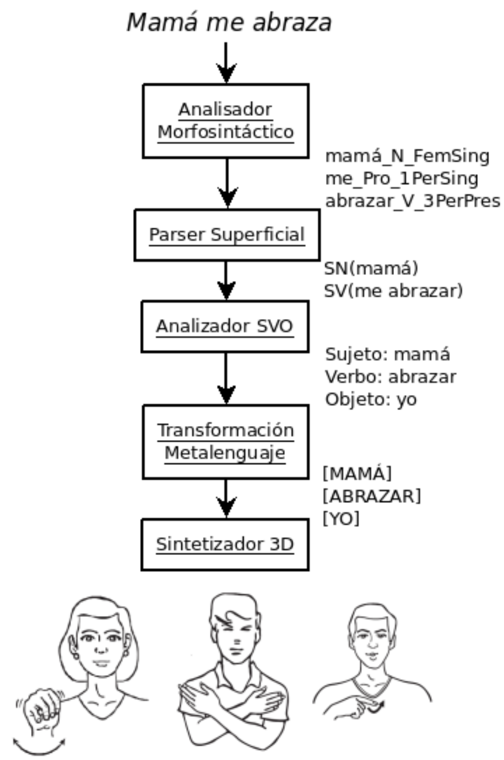


Fig. 6. Transformation of the Sentence "Mommy Hugs me" through the Modules of our Proposal. Own Source and Images of Signs Extracted from [4].

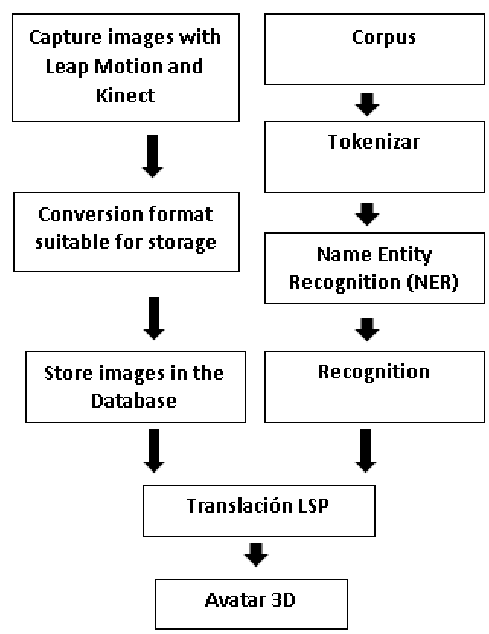


Fig. 7. LSP Translation Process using a 3D Virtual Assistant Own Source.

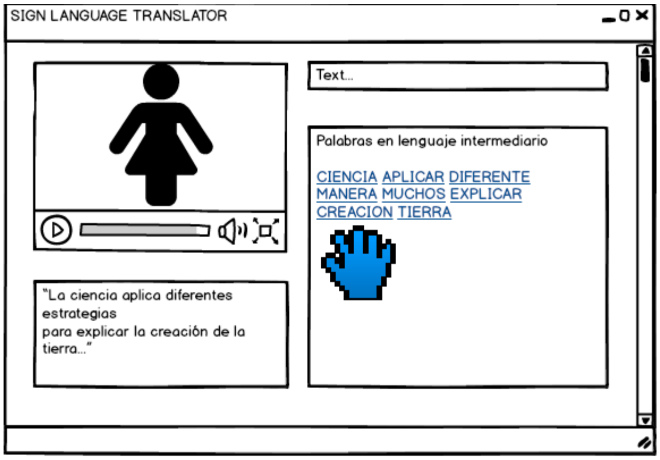


Fig. 8. Translation Page Mockup to LSP using a 3D Virtual Assistant Own Source.

The operation of the system is expected to be similar to that shown in the following Fig. 8, mockup was created in Balsamiq Mockup

a) Results and Limitations: The system is in the testing stage because it continues to improve to add new signs, in the process of creating the dictionary was discovered that there are signs not digitized in the guide of the LSP, because they are not recognized as official signs, until the presentation of this document in January-2020, the system recognizes words from a limited corpus of encyclopedia texts for children, due to the non-existence of all words expressed in an encyclopedia, is that the database of signs was limited, for example:

"La ciencia aplica diferentes estrategias para explicar la creación de la tierra"

Translated

[CIENCIA] [APLICAR] [DIFERENTE] [MANERA] [MUCHOS] [EXPLICAR] [CREACION] [TIERRA]

Where it is observed that the word *ESTRATEGIA* has no specific sign, due to this there is a change in the translation to the intermediate language for the use of signs.

VII. Analysis of Results

At the end of the investigation some results were obtained such as:

The proposed architecture is sustainable to allow the virtual assistant to generate well-made virtual signs. Thus, defining the correct meaning of what is written in the texts.

The translation allows a better understanding of the context of the sentences, changing words without meaning in signs to words or prayer with explanatory signs of the main meaning.

VIII. Conclusions and Future Works

After conducting research on the transformation of educational texts from natural language to sign language, the following conclusions are presented.

Creation of new material for the LSP community library with information for future research.

With this work we are collaborating to the social inclusion of deaf people, allowing them to improve their educational quality.

This research will serve for further extensions in handling a greater amount of corpus, in various mother tongues of different countries.

Acknowledgment

We would like to thanks to the Project funded by UNSA-INVESTIGA with contract IBAIB 08-2019 UNSA.

References

[1] INEI, "Primera encuesta nacional especializada sobre discapacidad," Instituto Nacional de Estadıstica e Informatica-Consejo Nacional para la Integración de la Persona con Discapacidad CONADIS, 2014.

[2] E. Gonzales and R. Guizado, "La interpretaci´on a lsp del himno nacional y del soneto clxvi de góngora," Academia Peruana de la Lengua, vol. 14, 2018.

[3] L. Peluso, "Traducción entre español escrito y lengua de señas uruguaya videograbada: un nuevo desafıo," Cad. Trad Florianopolis, no. 35, pp. 479–504, 2015.

[4] M. de Cultura Peruano, "Lengua de señas peruana : guıa para el aprendizaje de la lengua de señas peruana, vocabulario básico," Academia Peruana de la Lengua, vol. 1, 2015.

[5] Maldonado, "Trabajo y discapacidad en el perú," Mercado laboral,polıticas públicas e inclusión social de las personas con discapacidad, vol. 14, 2004.

[6] L. C. C. GUTIERREZ and A. G. R. ROSAS, "Presencia de los modelos curriculares en el diseño curricular regional de arequipa," 2015.

[7] M. de Educacion, "Curriculo nacional, educación basica," 2016.

[8] Universidad Nacional de San Agustín de Arequipa(UNSA),"Estudiantes con sordera estudian en la UNSA", 2019, Arequipa - Perú.

[9] E. Sosa, "Procesamiento del lenguaje natural: revisión del estado actual, bases teóricas y aplicaciones," Revista Nacional científica y profesional, vol. 1, 1997.

[10] V. Sandoval and R. Kenny, "Sistema intérprete de lenguaje alternativo para mejorar la comunicación de las personas sordas en la asociación de sordos de la libertad," association for Computational Linguistics, Stroudsburg, 2015.

[11] Johan Serrato Romero y Mario Chacón Rivas, "Traductor de documentos y texto a lengua de señas" , 2015, TEC Digital, Instituto Tecnológico de Costa Rica.

[12] "Lengua de señas peruana : guía para el aprendizaje de la lengua de señas peruana, vocabulario básico. 2015.

[13] Juan Carlos Quinche Curtidor, "Dispositivos de captura de movimiento (Kinect), para la navegación de experiencias formativas en ambientes virtuales 3D", Tesis de maestría Tecnológico Monterrey, 2015.

[14] Eric Estévez Rodríguez, "Captura y postprocesado de modelos 3D con Kinect", Universitat Politècnica de Catalunya, 2017.

[15] OH Diaz Jara, "Utilidad de la reconstrucción virtual en escenarios criminales de muerte por proyectil de arma de fuego usando el software iclone en el Perú 2015", Peru, 2015.

[16] P Alvarez, M Alejandro, "Desarrollo de un juego didáctico mediante sensores Leap Motion para estimular el aprendizaje del lenguaje básico de señas ecuatoriano", Ecuador, 2018.

[17] Thomas Tracey, Language Translation with RNNs Build a recurrent neural network that translates English to French, Aug 29, 2018.

[18] N. Raziq and S. Latif, "Pakistan sign language recognition and translation system using leap motion device,"International Conference on P2P, Parallel, Grid, Cloudand Internet Computing, vol. 9th, 2017.

[19] A. Gelbukh and G. Sidorov, "Procesamiento autom´atico del espa˜nol con enfoque en recursos l´exicos grandes,"Direcci´on de Publicaciones del IPN, M´exico, 2010.

[20] M. L. Palomar, M. M. A., and F. A., "Introducción al procesamiento de lenguaje natural,"Servicio de publicaciones de la Universidad de Alicante, Alicante , Espa˜na,1999.

[21] W. A. Woods, "Progress in natural language understanding: An application to lunar geology. en proceedings of the june 4-8, 1973, national computer conference ande xposition,"AFIPS, p. 441–450, 1973.

[22] O. Pichardo-Lagunas, B. M.-S. Luis Partida-Terron, A. Alvear-Gallegos, and R.Serrano-Ole, "System for direct translation from spanish into lsm with markedrules,"Instituto Polit´ecnico Nacional, Unidad Profesional Interdisciplinaria en Ingenieria y Tecnologıas Avanzadas, Ciudad de M´exico, M´exico, 2016.

[23] A. Gelbukh and G. Sidorov, "Procesamiento automatico del espanol con enfoque en recursos lexicos grandes,"Direcci´on de Publicaciones del IPN, M´exico, 2010.

[24] S. Abney, "Part-of-speech taggering and partial parsing. in corpus based methods in languaje and speech,"kluwer Academic publishers, pp. 117–136, 1997.

[25] M. F. Xavier and M. H. M. Mateus, "Dicionario de terminos linguısticos,"Edicion Cosmos, p. 441–450, 1990-1992.

[26] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard y David McClosky. 2014. El kit de herramientas de procesamiento del lenguaje natural de Stanford CoreNLP en las actas de la 52ª reunión anual de la Asociación de Lingüística Computacional: Demostraciones del sistema , págs. 55-60.

[27] Ševčenko, M. 2003. Online Presentation of an Upper Ontology. CTU Prague, Dept of Computer Science.

A Service-Oriented Architecture for Optimal Service Selection and Positioning in Extremely Large Crowds

Mohammad A.R. Abdeen
The Faculty of Computer and Information Systems
The Islamic University of Madinah, Madinah, Saudi Arabia

***Abstract*—The problem of managing large crowds has many aspects and has been reported in the literature. One of these aspects is the distribution of supplies such as food and water in especially when the targeted region is overcrowded. Some of the challenges is to plan the locations of food and water supply centres in such a way to achieve multiple objective functions such as the type of food and the shortest distance to customer. A practical example of this problem is the food distribution and food cart location in the region of Mena (also known as Tent City, in Saudi Arabia) during the yearly pilgrimage season. In this work, we propose a Service Oriented Architecture (SOA) for positioning services in the region of Mena (Mecca – Saudi Arabia) that covers an area of approximately 20 square kilometres during the pilgrimage season. The architecture proposes an optimal service selection as well as a mobile food cart positioning algorithm based on client pre-set profiles to achieve multiple objective functions for the clients as well as the service providers. Some of these objective functions are the least waiting time to be served, the shortest distance to service, the lowest cost, and the maximum profit for the service provider.**

Keywords—Large crowd management; Service-Oriented Architecture; multi-objective optimization; Hajj; Mena; WSDL

I. Introduction

Each year, millions of pilgrims travel to Mecca from various places/countries all over the world to perform their once-in-a-life-time duty of pilgrimage (Hajj). The trip lasts for at least five days in which pilgrims move from a place to another within an area of approximately 50 square kilometers (12x4 Km). Managing such a large number of people, that reaches two million, is a significant challenge to the event organizers. A variety of services are offered to those pilgrims including, the supply of food and water, accommodations, healthcare, guidance, and many others. Facilitating such services for such a large number of people and in such a relatively small area with dense population is a major undertaking. One of the regions that pilgrims pass through and stay for three days is called the region of Mena, a.k.a. the "Tent City". This place is located in the open desert and it is economically infeasible to build infrastructure to house pilgrims and services for just three days in the whole year. The pilgrimage's organizers have resorted to the idea of pre-prepared tents that are easily installed and moved around. Fig. 1 shows the region of Mena [1].

With today's technology and the vast availability of mobile devices, there comes an opportunity to organize such events with better efficiency and convenience for both the pilgrims and the organizers.

Food and water distribution in the region of Mena (in Mecca) is usually done by trucks carrying various food supplies [2]. Also, the authorities have recently depended on water containers for water distribution in the region of Muzdalifa (few kilometres away from Mena) [3]. A dire question is where to locate those food mobile trucks and water supply containers so that an optimal service is provided to minimize the effort and time for pilgrims.

This work proposes a Service Oriented Architecture (SOA) that facilitates positioning mobile service carts for the purpose of mutual benefit for both the consumers (pilgrims) and the service providers (such as food, hairdressing, telephone services, etc.). Since the area of Mena has some permanent but insufficient installations for such service provisioning, this architecture proposes the smart use of mobile service carts to provide the required services. The architecture uses the GPS location of clients in addition to their pre-set profile in which they specify the type of service required. The system proposes to clients the most convenient service (with shortest distance, and least service times). The architecture also uses the customer information to dynamically locate mobile carts to achieve the same goals based on clustering criteria that combines the customer's and the service provider's own criteria.

This paper is organized as follows: Section 2 presents a background of the topic and the definition of the main technical terms. Section 3 presents the proposed SOA. Section 4 shown the implemented prototype of the system. Conclusions and future work are presented in Section 5.

Fig. 1. The Region of Mena – Mecca (The Tent City).

II. BACKGROUND

Large crowd management has been addressed in several works in the literature [4][5][6].

In [4], the authors presented a broad review on various crowd management and monitoring technologies that are based on vision (CCTV), Wireless/RF and Web/Social media technologies. It presented two of the most known examples of the crowd management, viz., the Hajj (pilgrimage) and the Kumbh Mela (in India). They addressed the events form the security view point and to employ the latest technology to avert disastrous situations.

The authors in [5] presented a Near Field Communication (NFC) based architecture for providing various services in large crowd situations. The architecture relied on mobile phones that are NFC enabled. The authors presented the Hajj (pilgrimage) as a case study. They considered services such as medical emergencies, pilgrim identification, and lost pilgrim help. They implemented a proof-of-concept on a Samsung smart phone that runs Google Android OS. A Microsoft SQL server is used in the backend.

In [6], the authors presented an architecture for providing a specific, yet important, service for individuals in large crowds. They presented an in-memory architecture for locating, tracking, and guiding astray pilgrims during Hajj season in Mecca. The simulation results showed that the architecture scales well with the number of pilgrims in the system.

In [7] the authors presented an architecture to manage large crowds in the region of Mecca during the pilgrimage (Hajj) season. The architecture addressed managing the large crowds including security issues and using RFID to track pilgrims. However, they did not take into account the provision of food and other services.

A. *Servie Oriented Architecture*

Service oriented architectures is a software approach to developing efficient general-purpose distributed applications from specialized components called services. This approach significantly improves productivity and reusability. The SOA has many useful applications in the banking [8], healthcare [9], mobile-learning (m-learning) [10], automotive [11] and many other sectors.

Services are the central part of the architecture which are components with well-defined interface which are raceable from any location in the network. Each service defines a contract with which the service is accessed. The architecture has essentially three components, the service provider, the service requester and the service registry. Fig. 2 shows the components of a SOA.

Some standards are an integral part of the SOA, such as the WSDL, the UDDI, and the SOAP. In the following sections these topics are briefly covered.

B. *Web-Service Description Lanuage*

Services need to identify themselves to the service registry. The Web Service Description Language (WSDL) is an xml-based language that enables service providers to describe their offered service in a machine-readable format. It includes definitions of how the service is called, what parameters it takes, and what data structures it returns. A WSDL file is an XML-based file and is where the service description is stored. It has several components including the service definition, the data types, the messages, the port type, the binding, the ports, and the services. The service definition in the form of a WSDL file is published to a service repository [12].

C. *Universal Discovery and Description Integration*

An important component of the SOA is the service registry. This component provides for the ability to search and discover the registered services. Service requesters are the primary customer for such a component. The Universal Discovery and Description Integration (UDDI) is a technology based on XML markup language and was developed by a consortium of more than 300 businesses and technology leaders to enable companies and/or applications to quickly find web services. The UDDI registry is essentially a directory and consists of three sub-directories:

- The white-pages: where organizations and service providers list their names, phone numbers, and address.
- The yellow-pages: where more detailed information such as business classification according to the north-American standard.
- The green-pages: where information about the business process is included such as shipping, billing, and purchasing methods.

D. *The Simple Object Access Protocol*

A standard protocol is needed to wrap all messages that are being exchanged between the components of the SOA such as messages between the service provider and the registry service, the service requester and the registry service and finally the service requester and the service provider. The Simple Object Access Protocol (SOAP) just provides that. It is a standard XML-based wrapper protocol that wraps all communication for a SOA application. A typical SOAP message consists of an envelope that contains a header and a message body. The header contains information on how the message is to be processed including routing and authentication information. The message body, however, contains the actual message to be processed and delivered [13].

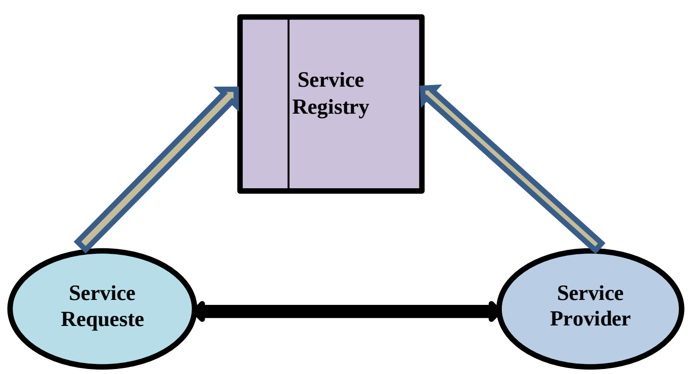


Fig. 2. The Service-Oriented Architecture.

(*sponsors*).

III. The SOA Architecture

In this section we describe the overall high-level SOA architecture for the proposed system.

The presented architecture consists of three main components similar to the standard SOA, as shown in Fig. 3. However, the type of component and its behavior is what enables this architecture to serve its purpose. The service requester is the client requesting various services, such as food services, telephone service, hairdressing service, transportation services. The service provider is the entity that provides the variety services for pilgrims and visitors as previously mentioned. The registry is the entity where all services have to register with prior to being acknowledges in the system and to enable the search and discovery by the client.

A. The Clients (Service Requester)

The client is the service requester and it runs a mobile application on a mobile device. Initially the client selects through the mobile App the specific service from a radio-button selection menu. The client also selects his/her preferences such as type of food for example (ethnic food, fast food, Veggie) or mobile service provider preferred (Vodaphone, Bell, Orange Telecom). This data is sent to the service registry for recommendations.

B. The Registry

The registry is an entity (a process) that runs on a separate machine that keeps information about the services that are registered with it. In this architecture, the registry keeps information about the food service providers, the phone service providers, and other service providers. A service providers registers to the registry by sending its information in a WSDL message. The registration information includes, name of service provider, type of service offered, return value type (a data structure in this case).

C. The Services

There are various types of services supported by the system, the food services, the mobile phone service, the hair dressing service, and the transportation service. These services register initially to the service registry and provide information regarding the provided service. This information includes the following items:

- Service Type (Food, Hairdressing, telephone, …)
- Service GPS Location
- Service cost.
- Available Deals/discounts
- Average time to serve
- Delivery available (Yes/No)

Service providers are in the form of mobile carts and are initially located as per a “first guess” provided by the service optimizer component. The service providers move later to a possible future location as per the recommendation of the optimizer which utilizes an algorithm based on centroid calculations.

D. The Service Optimizer

In this architecture an extra service is proposed that has a central role for the system operation. This service is provided by the facilitator or the framework. It receives location information from service requester as well as the optimization criteria as per service requester (whether the requester wants to optimize cost, delivery time, or distance) and it replies with the service provider reference that achieves the requested optimization criteria.

Fig. 4 depicts the sequence diagram of the proposed architecture and the interaction of its components.

As shown in the sequence diagram, the client selects the type of service from a mobile App on the client smart phone. The client is provided with further detailed choices according to the type of service selected. As an example, if the client selects food service, a list of other options such as type of food (ethnic, fast food, special diet – Veggie, Gluten free) is displayed. Other options such as pick-up or delivery options are provided. This request is forwarded to the service registry which keeps a list of all registered service providers. Supported service providers are required to register their service with the registry and provide service details such as service type, service location, average waiting time, and whether or not there is a delivery option. Upon receiving a client service request, the registry service replies back with a list of possible candidates that provide the requested service. The client then retrieves the references of the received candidates and calls the optimizer service after obtaining its reference and sends the records of the service candidates.

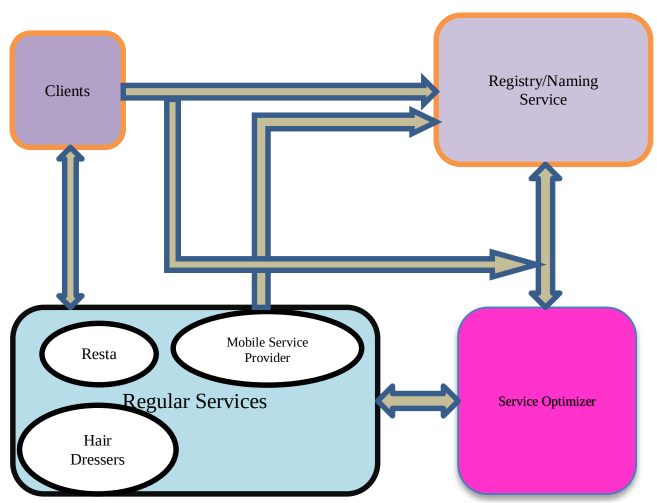


Fig. 3. The Proposed SOA for Food and Service Distribution in the Region of Mena, in Mecca During Pilgrimage.

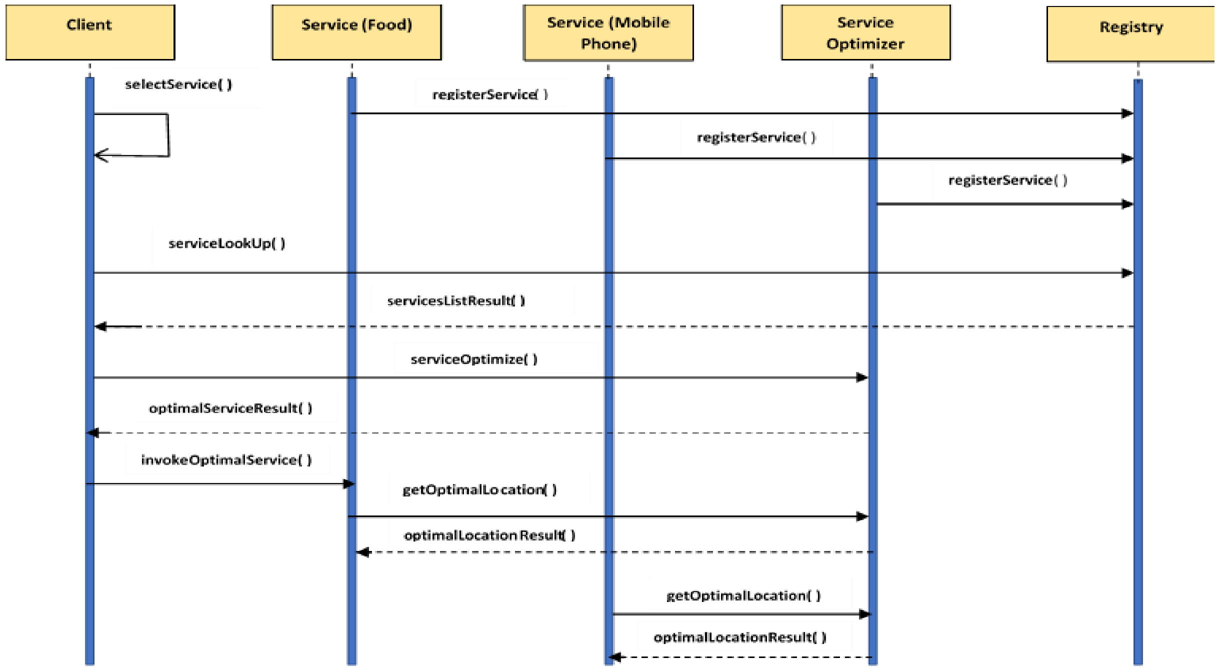


Fig. 4. The Service Distribution Sequence Diagram.

1) Thin client–This architecture suits the utilization of thin clients where only data presentation is handled by the client and the actual computation and optimization algorithms are implemented on the server side. This architecture suits well the majority of the mobile devices and to maintain the availability of this type of service on as many devices as possible to suit the diversity of the pilgrims and their mobile devices of choice.

2) Fat client–The architecture can also support fat client in which client device handles both the data presentation as well as the data processing algorithms. In this case, the client device downloads the mobile App as well as a plugin that implements the optimization algorithms. In this situation, the client needs not to communicate back-and-forth with the service optimizer component.

IV. Prototype Implementation

We have implemented a prototype of the proposed architecture using Android Studio [14] as the mobile App development environment and the Visual Studio [15] environment. Fig. 5 shows the initial menus of the developed mobile App. The first menu shows the menu items at the client side which lists the available service to the client such as food, styling, telephone, and transportation. If a client, for example, selects the food option, a submenu appears showing the subitems (such as Ethnic, Veggie, and Fast food). Upon the client pressing the "Go" button, a SOAP message is sent to the registry service requesting the availability of the selected item. The registry responds with a reply SOAP message showing all options in the required class and the reference to every service. These services have previously been registered to the registry with an WSDL message. The client uses the reference information to directly contact the service providers to obtain their relevant information (as shown in the right graph of Fig. 5).

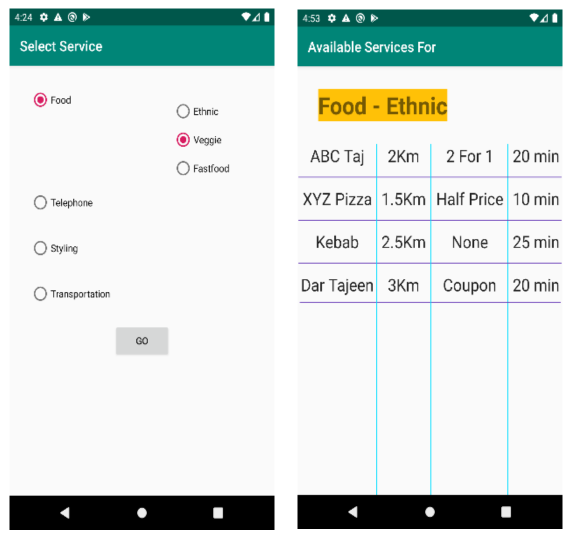


Fig. 5. The Mobile App Showing the Service Selection and Detailed Service Response.

Clients then can opt to optimize the list of services provided by the service providers. Clients can call the Optimze_Service() service and provide their optimality criteria. This service sends back the response to the client with the optimal list. The code Listing 1 shows the Optimize_Service() web service.

On another hand, this system provides an optimal location service to service providers. In other words, a service provider can request that it sent the location where the majority of customers reside. As an example, if a given food cart offers ethnics food of a specific ethnicity (e.g. Indian) then the service provider can request to be sent the optimal location that provides the minimum average distance to the majority of those clients. Fig. 6 shows a schematic of the optimal position calculation based on the concept of centroid. If the clients are scattered across a given area, then the optimal location of a service cart is the centroid of the service requester as shown by Eqn. (1).

$$Centroid = \frac{1}{n}\sum_{i=1}^{n} d_i \quad (1)$$

where d_i is the actual distance between an initial centroid location and a given client, and n is the number of clients belonging to the class of service (such as ethnic food). Listing 2 shows the code for **Optimize_Location()** web service that is used by the service providers.

```
enum OptimalVals {proximity,
   waitingTime, Price, Deal};
class GPS{int x; int y};
class Criteria{
   GPS clientGPS;
   OptimalVals[ ] List = new
OptiamlVals[5];
}
class cItem{
   string name;
   float  dist;
   float  discount;
   int    waitM; //waiting time in
minutes
}

//this function is a service called by
//a service requester to obtain best
// service according to a priority
// list he/she provides.
cItem Optimize_Service(Criteria
OptCria){
         Console.WriteLine("Optimizing
                Service");
       cItem CI = new cItem();
       return CI;
}
```

Listing 1 Sample Code for the Service Optimization Web Service.

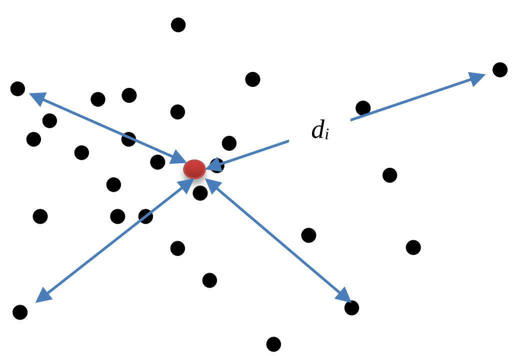


Fig. 6. The Client Centroid Calculation.

```
enum ServiceTypes {Food, Styling,
Telephone, Transportation};

enum ServiceSubTypes{ Ethnic, Veggie,
Fastfood, HairDress, Vodaphone, Orange,
STC, Taxi, Bike, Bus};

class ServiceType{

   ServiceTypes type;

   ServiceSubTypes  stype;

}

GPS Optimize_Location(ServiceType sT){

   Console.WriteLine("Optimizing
Location");

   return gps;

}
```

Listing 2 Showing the Optimize_Location(..) Web Service.

V. Conclusions and Future Work

In this paper we presented an SOA architecture for service provisioning and optimal service location in events with extremely large crowds. We considered a case study of the Hajj (pilgrimage) and more specifically the region of Mena where fixed installations and facility are not economic justifiable and mobile service carts are more appropriate. A prototype of the architecture was implemented as a mobile App and services are implemented with Android Studio and Microsoft Visual Studio. One important component of the system is the service optimizer. Upon receiving a client request for a service and based on the client optimal criteria, the system returns a sorted list of the recommendations including the average waiting time to get the service. In addition, the services can solicit the optimal location from the service optimizer to be convenient for the majority of the service requester. Optimization of the service location includes commuting times, and proximity of a special-type customers (such as when selling ethnic food).

As a future work, scalability of this architecture is an important issue. The architecture should be able to support hundreds of thousands of users simultaneously. A service registry could become a bottle neck due to surges in requests during peak times. In addition, security is another concern. Since service are public, it is then available for productive and nonproductive users. It is of equal importance to prepare for hundreds of thousands of legitimate users as well as for other users that might misuse the service.

Acknowledgements

The author would like to thank Dr. Khaled Bashir for his help with the mobile App in this work.

REFERENCES

[1] The International Business Times Available From https://www.ibtimes.co.uk/hajj-2014-photos-vast-crowds-over-two-million-muslim-pilgrims-gather-mecca-1468734 [last accessed: December 23, 2019].

[2] Al-Watan Newspaper , Available From https://www.alwatan.com.sa/ article/67972/الاقتصاد/يقدم-أكبر-مائدة-طعام-وشراب-لملايين-الحجاج-في-مدينة [Accessed: December 2019].

[3] Saudi Press Agency (SPA), Available From, https://www.spa.gov.sa /1957009 [Accessed: December 2019].

[4] D. Sharma, Amol P. Bhondekar, A. K. Shukla, and C. Ghanshyam. "A review on technological advancements in crowd management." Journal of Ambient Intelligence and Humanized Computing 9, no. 3 (2018): 485-495.

[5] M. A. Mohandes,. "Mobile technology for socio-religious events: a case study of NFC technology." IEEE Technology and Society Magazine 34, no. 1, 2015, pp. 73-79.

[6] M. A.R. Abdeen and A. Taleb. "A Real-Time Algorithm for Tracking Astray Pilgrim based on in-Memory Data Structures." International Journal of Computer Science and Applications 9, no. 10, 2018, pp. 467-474.

[7] M. Yamin and M.A. Albugami, “An architecture for improving Hajj management”, In: Liu K, Gulliver SR, Li W, Yu C (eds) Service science and knowledge innovation ICISO 2014 IFIP advances in information and communication technology, vol 426. Springer, Berlin, 2014.

[8] J. DiMare, and R. S. Ma. "Service oriented architecture Revolutionizing today's banking systems.", IBM Global Business Services, 2009, [Available From: https://www.coursehero.com/file/32455299/SOApdf].

[9] K. Avila P. Sanmartin, D. Jabba, and M. Jimeno. "Applications Based on Service-Oriented Architecture (SOA) in the Field of Home Healthcare." Sensors 17, no. 8 , 2017.

[10] M. J. C. Guerrero, , M. González, M. A. Forment, and F. J. García-Peñalvo. "Applications of Service Oriented Architecture for the Integration of LMS and m-Learning Applications." In WEBIST, 2009, pp. 54-59.

[11] M. Wagner, A. Meroth, and D. Zöbel. "Developing self-adaptive automotive systems." Design Automation for Embedded Systems 18, no. 3-4, 2014, pp. 199-221.

[12] Web Service Description Language Version 2.0, Available From https://www.w3.org/TR/2007/REC-wsdl20-20070626/ [Accessed: December, 23, 2019].

[13] S. Graham, G. Daniels, D. Davis, Y. Nakamura, S. Simeonov, P. Brittenham, P. Fremantle, D. Koenig, and C. Zentner, Building Web services with Java: making sense of XML, SOAP, WSDL, and UDDI, SAMS publishing, 2004.

[14] The Android Studio, Available From, https://developer.android.com/ studio [Accessed: December 2019].

[15] The Microsoft Visual Studio Available From https://visualstudio. microsoft.com [Accessed: December 2019].

EDUGXQ: User Experience Instrument for Educational Games' Evaluation

Vanisri Nagalingam[1], Roslina Ibrahim[2], Rasimah Che Mohd Yusoff[3]
Advanced Informatics School, University Technology Malaysia
Kuala Lumpur, Malaysia

***Abstract*—A significant increase in research on educational computer games in recent years has proven that the demand for educational games has increased as well. However, production of incompatible educational games not only cost wastage of money but also energy and time for game designers and game developers. To produce a suitable educational game, it is important to understand the user's need as well as the educational need. Therefore, this study aims to develop a User Experience (UX) framework for educational games (EDUGX) based on UX elements and psychometrically validate a new instrument, EDUGX questionnaire (EDUGXQ) that is appropriate to evaluate educational games. Based on literature review, six main UX elements were identified which are Flow, Immersion, Player Context, Game Usability, Game System and Learnability to construct the framework. In this paper, we first discussed the development process of EDUGX framework followed by EDUGXQ. This study will also review and discuss several UX questionnaires for educational games in UX design evaluation which at the same time supports the framework's elements to develop the EDUGXQ.**

Keywords—User Experience (UX); framework; psychometrically; educational games; educational games' evaluation

I. Introduction

The era of the gaming world has grown extensively as the recent internet trend report shows that the number of gamers from 1995 to 2017 has reached tremendous height, with an increase of 2.6 billion gamers in 2017 versus 100 million in 1995 with a global gaming revenue of $100 billion in 2016 [1]. This sudden rise of computer games usage as a favorite past time activity has increased the attention in using educational games to help in skill acquisition, behavior change, attitude and learning[2] and created a significant increase in research on educational computer games in recent years. Problems such as lack of game content or learning content often occur when designing a computer game with learning purposes [3]. Nevertheless, there is lack of agreement among game aficionados in deciding what features are essentials to the creation of successful or good games [4].

Since it's still a new knowledge and in its inarguable phase, the game developer considered the field of educational game as new and risky. Moreover, the designing and developing of educational computer game would require a large budget, resources and special skills. According to [5], there are also needs to have an appropriate level of educational accuracy to suit the learning subject and the classroom environment. Poor and irrelevant design of a product may resulted in unwanted or wasted production, money and energy. This clearly shows that educational games production can be a difficult process.

To understand the complicated process of developing educational games, the conception of evaluation is required. An evaluation is required as it allows describing a framework which will be verified by third parties in different perspectives and the results of the evaluation can assist the developer or designer to pinpoint the errors or negative points of the game. Even with a growing body of evidence on the efficacy of computer games for learning, evaluation is often incomplete, biased, poorly designed, if not absent [6]. In addition, developing and evaluating games for specific purposes whether for behavioral change or learning is a very challenging enterprise.

Currently, most of the evaluation or testing of computer games during the development stage are done with technical testing such as bug testing and usability testing. This testing ensures that computer games can perform as per their technical requirement and game functionality. However, when a player uses computer games, it becomes more than technical or functional aspects. Game which consist of the non-instrumental qualities are more important to be evaluated compared to other digital systems [7]. Non-instrumental qualities can be defined as quality features of an interactive system that address user's needs that go beyond goals, tasks and their efficient achievement [8]. These qualities involved the user's personal preferences and emotions. Therefore, for this purpose an UX evaluation is needed to identify the user's preferences and responses [9] as each of the player's experience is unique.

This unique experience can be evaluated if the right UX elements are measured [9]. Hence, an UX evaluation framework is especially important for predicting, understanding and reasoning about procedures of UX [10], which indirectly help the game designers to evaluate their games. By producing a framework on the UX evaluation for educational games, it helps the future educational games designers to understand the needs of their users and indirectly helps in developing productive educational games with cost-effective learning design and game development.

Thus, in this study, the key elements for UX evaluation will be identified and tested to identify if the games contain the elements to become a successful game. User Experience Evaluation for Educational Games (EDUGX) framework was developed to support UX evaluation for educational games followed by the development of the gaming scale EDUGX questionnaire (EDUGXQ) that is based on the EDUGX framework. This section includes the background of the present research, research approach and its aim. Section 2 discusses the related works in the process of EDUGX framework followed by section 3 which focuses on the construction of EDUGXQ and discusses the expert review on the proposed EDUGX and EDGXQ. Followed by the discussion on the findings. The last section is the conclusions and future works of this paper.

II. Related Works

User experience is focused on the interactions between products and people, and the experience that resulted in certain contexts of use [15]. This experience can be evaluated if suitable elements are measured. To identify the suitable elements of UX evaluation for educational games, the previous models/frameworks of UX are reviewed. The literature has indicated that many elements can be used for computer games evaluation. However, only six main elements are used by most of the researchers for UX evaluation as presented in Table 1.

Based on the six elements reviewed earlier, all of these elements are identified as suitable for UX evaluation for educational games. Flow, immersion, game usability, game function/system, player context and learnability are added to the proposed conceptual framework. These six elements support each other to find the right balance of the user experience and educational games. The suitable sub-elements of each of the elements are also identified based on the earlier review. The following are the elements and their sub-elements.

A. Flow

The template Flow state engages a person in a goal-directed, complex challenge merely for the pleasure of dealing with the challenge [11]. As for flow, goal clarity is an important sub-element to achieve the goal of the educational game. Meanwhile, sub-element feedback enables players to understand their status of the game in real time. Sub-element challenge ensures that the players do not lose their interest of the game by providing an appropriate level of challenge according to the player's skill level. Another sub-element under flow is concentration which ensures that the game quickly grabs the players' attention and maintain their focus throughout the game [12]. Additionally, sense of control towards the game is also important to create a fun gameplay experience [13].

Flow is a state that needs to be sustained with supporting sub-elements such as sense of control, sufficient feedback and clear goals to ensure the players are fully immersed in the game's challenges [14]. In educational games, concentration is important to ensure that the players receive the knowledge behind all the gameplay. Thus, clear goals, feedback, playability, control, challenge and concentration are listed under Flow element for this study.

B. Immersion

Immersion is a state that makes players believe that they are in the game content and directly involved in the game world [15]. In addition, [16] have employed grounded theory to explore game immersion and suggested that immersion is involved in three stages: engagement, engrossment, and total immersion, respectively using Game immersion questionnaires (GIQ).

Most immersive gaming experiences were of the engrossing or engaging variety with total immersion reserved for the most intense periods and even restricted to a short period within a longer playing session [17]. The lowest level of involvement is engagement. Engagement is dependent on the gamer's willingness to invest time, attention and effort in the game meanwhile the engrossment level affect players by involving player's emotions. In the last stage, total immersion will cut off the player from reality into the game world [18].

A recent study involving immersion, engagement and flow in game-based learning shows that both the challenge of the game and being skilled in the game had a positive effect on both being engaged and immersed in the game. In this case, challenge which is a sub-element of flow acts as a strong predictor of learning outcomes [19]. However, the study shows that immersion did not have a significant relationship with perceived learning and suggested that different types of game that involve sensory or imaginative immersion can be tested besides challenge-based immersion game.

C. Game Usability

[20] defined the "usability" as the possibility and the ability to have contact with a product in terms of satisfaction, efficiency, learnability, errors and memorability for older adults. Usability plays a significant role to make sure that the user can accomplish the goal of the product effectively and efficiently at an optimal time [21]. One of the objectives of ISO 9241-11 is to ensure that the satisfaction component of usability contains features of user experience [22]. According to [23], usability can be broken down into five sub-elements; Operability, Understandability, Learnability, Attractiveness, and Satisfaction.

When an immersive state is achieved, it could help to overcome other usability issues [24]. Besides that, the usability scores may also be positively or negatively affected when a game delivers experiences of immersion and flow [25]. Usability also supports the UX evaluation as good usability, a useful artefact and an engaging task (challenges that the game provides) create conditions for a good educational experience [26]. As for usability, the sub-elements of attractiveness, understandability, satisfaction and operability are included as all of these sub-elements play an important role in supporting the main element accordingly.

D. Game System

Based on the models proposed by [27-29], the game experience or player experience involved three layers of interaction: game system, player and context. The quality of the game system is important as player's experience is shaped when a user interacts with the game system. The game system plays an important part to ensure that the game can be played without any hiccup. A hiccup in gameplay can ruin the concentration of a player thus indirectly affect the immersion and flow experiences towards a game.

Some of the methodologies for assessing game system are soak testing, open beta-testing, localization testing, unit testing, compatibility testing, regression testing, bug tracking, stress testing and gameplay metrics [27]. In a learning environment, a system should be able to support both sides of communication (teacher and students). Therefore, the engine should be able to guide the process and command the games to activate their adaptation mechanisms to fit certain requirements every time they are run [30]. The device and functional level of a game system must be tested by each of the game development team to ensure it is working correctly.

E. Player Context

Player context can be explained as an interaction between the contextual gameplay experience formed and a player in a given temporal, social, spatial context [27]. As for the player context, many features can be measured to evaluate the user experience. User background, time, and culture are some of the important aspects to be considered. Culture reflects the environment that a player resides thus game related to the player's culture or environment can be more reachable. User background can also help the game designer to get some ideas of what the players prefer in a game and the level of the player's gaming skills. A game that relates more to the reality of the players can connect easily with the player. Thus, player's background needs to be identified.

As for time, it is unpredictable how the players will react to a game after a certain period. However, for an educational game, time cannot be considered as reliable since a subject used for the learning process will be acquired for a certain short period only. For example, for computer science students who use an educational game to learn a basic programming language, they will only use that game for that particular period or in certain semester. For the following semester, a new batch of students will join the class and use the same educational game. Besides, different educational game will have different learning objectives (learning content) that will only be used for a limited time. Therefore, sub-element of time can be ignored by this player content element.

F. Learnability

The main idea of a game is to have fun meanwhile learning is always hard, and making people learn through games can ruin the fun [31]. Thus, the right balance of learning and fun should always be the important principles of educational games. If the fun elements control over the learning goals, then the real objective of educational games will fail. Moreover, the aim of an educational game should be related to the learning goals of the game [26].

Without the right amount of educational content in the game, the real objective of the educational game will not be achieved. Hence, it is important to ensure that the learnability element is included in the user experience evaluation for educational games. The sub-elements of knowledge improvement, learning goals and learning content are included in learnability. Knowledge improvement is one of the key points of educational games as the objective of educational games is to ensure that knowledge of a particular subject is improved. It is considered as aspects of an enjoyable experience and an important criterion in evaluating educational games. Learning content and learning goals also support learnability to achieve the goals.

III. Methodology

Based on the review, the proposed EDUGX is modelled. Table 1 shows the list of all sub elements and the sources to model the EDUGX framework.

Sub elements from EDUGX framework will be used to develop the questionnaire tool (EDUGXQ) for validation and an initial expert validation will be carried out on the framework and questionnaire before the data collection process.

A. Identification Of EDUGX Framework

This UX framework was developed based on six elements which are Flow, Immersion, Player Context, Game System, Game Usability and Learnability. Questionnaires used by previous researchers for evaluation of user experience and the elements involved in each of the questionnaires are listed in Appendix A. A total of seven instruments that are used for educational games evaluation are reviewed. For each instrument, the elements involved are identified and cross-checked with the elements in the proposed EDUGX.

TABLE. I. Summary of all Elements and Sub-Elements for EDUGX

Elements	Sub Elements	References/Sources
Flow	Challenge	[14, 19, 23, 26, 32-35]
	Clear goals	[23, 35, 36]
	Playability	[15, 31, 37]
	Feedback	[35, 37, 38]
	Control	[33, 34, 37]
	Concentration	[26]
Immersion	Engagement	[16-18]
	Engrossment	[16-18]
	Total Immersion	[16-18]
Game System	Devices	[28-30, 39]
	Function	[28-30, 39]
Game Usability	Operability	[23]
	Understandability	[23]
	Satisfaction	[20, 23]
	Attraction	[23]
Player Context	User Environment	[27, 28, 38, 40]
	Prior Experience	[27, 28, 40]
Learnability	Knowledge Improvement	[9, 23]
	Learning Content	[9, 23, 35, 41]
	Learning Goals	[9, 23, 35]

Out of the seven instruments, four questionnaires included flow as their element. UGALCO and Gameflow questionnaires involve knowledge and learning experience as one of their elements. GIQ is a thorough questionnaire on Immersion elements such as Engagement, Engrossment and Total Immersion. In addition, this questionnaire needs to be seriously taken into consideration as immersion is one of the elements for EDUGX. Therefore, GIQ is considered as one of the important questionnaires to investigate. Another element that needs to be evaluated in EDUGX is the playability. Thus, PLEXQ is considered as one of the important questionnaires to be adapted into EDUGX questionnaire.

B. EDUGX Questionnaire (EDUGXQ) Development

Based on the review of the existing questionnaires, UGALCO, PLEXQ, GIQ, Game Engagement Questionnaire and Gameflow questionnaires are selected for the development of the EDUGX questionnaire since these questionnaires focused on UX in games. To develop the required questionnaire, the elements involved in the six main elements of EDUGX are picked from the selected questionnaires. As for the game system element which is focused on measurement done by the game development team, a few questions will be added to understand the level of functionality from the user perceptions [27, 29]. Once the elements needed for the EDUGX questionnaire development are identified, applicable questions are then selected. This is to avoid long or repeated questions.

Questionnaire related to the six elements on the proposed EDUGX framework was selected with a total of 99 questions. Closed questions with multiple tick boxes, yes/no choices and ranking with 5-point Likert scales were applied as per existing questionnaire for a quicker and easier approach. The selected questionnaire items for the proposed EDUGX questionnaire (EDUGXQ) is listed and EDUGXQ will be finalized once experts' reviews suggestion and comments are obtained.

C. Expert Review

In this section, the proposed EDUGX framework and EDUGX questionnaire are shown to the UX and game industry experts for validation. The framework and questionnaire development will then be finalized. The objective of the expert review is to avoid any big changes to the EDUGX framework or questionnaire after the data collection phase. Once the experts validate the framework and questionnaire, any changes or update will be done according to their suggestion.

An official email request was sent to a total of eight experts (five academicians and three game developers). From the eight experts, four academicians in the field of User Experience and one game developer from the game industry agreed to be in this research expert panel. Following that, their details were taken, and a set of expert review form was sent to them by mail and also face to face meeting, according to their preferences. This expert review form consists of five main sections:

a) Introduction: Include demographic and research work information.

b) Consent form: Experts can read, understand and agree to the reviewing process.

c) EDUGX Framework: Brief description of what is EDUGX and how it was developed.

d) EDUGXQ: The questionnaire was arranged according to the six elements with each item on a four-point scale (1-not relevant, 4-highly relevant).

e) Overall Comments/Suggestions: Experts are required to answer five questions on the framework and a questionnaire based on their opinion.

IV. Findings

Based on the result obtained from the experts, the validating process was done to make changes and discard any irrelevant items. The common comment among the experts was to change the negative statements to positive statements to avoid confusion since it's hard to evaluate negative statements. Other comments were, to make simpler and clearer questions/wordings.

Validity ensures that the questions being asked allows for valid inferences to be made. Since this study involves students and educational game, the type of validity is selected based on this category. There are four types of validity in educational research which are criterion-related validity, construct validity, content validity and face validity [42]. This study will use content validity to ensure that the items in the questionnaire that addressed each of EDUGX elements will be evaluated.

Content validity has been defined as the extent to which an instrument has an appropriate sample of factors for the questionnaire being considered [43]. The content validity index (CVI) was used to measure content validity, which is 'based on experts' rating of item relevance [44]. Based on the 4-point scale, the rating of 1 and 2 is considered as not agreed as relevant items and the rating of 3 and 4 is considered as agreed items to be relevant. Thus, an item level CVI (I-CVI) is used to calculate the degree of agreement among the expert panelists.

The I-CVI score was computed for each item by adding the number of experts who rated an item either 3 or 4 and dividing it by 5. This number shows the proportion of experts who agreed that the item is relevant. If all five experts rate an item as either 3 or 4 in relevance, the item CVI will be 1.00. According to [45], CVI value of 0.78 or higher from more than three experts can be considered as good content validity. Thus, for this study, the items with CVI greater than 0.8 were included in the final EDUGXQ. All of the items were listed to identify the CVI value as shown in Appendix B.

From a total of 99 questions, 24 questions were removed from the initial EDUGXQ and sentence structure of the remaining items were also restructured based on feedbacks received from the experts. The revised EDUGXQ contained 75 items of 5-point Likert scale with multiple choices options, categorized under five thematic domains namely (1) Flow; (2) Immersion; (3) Player Content; (4) Game Usability and (5) Learnability.

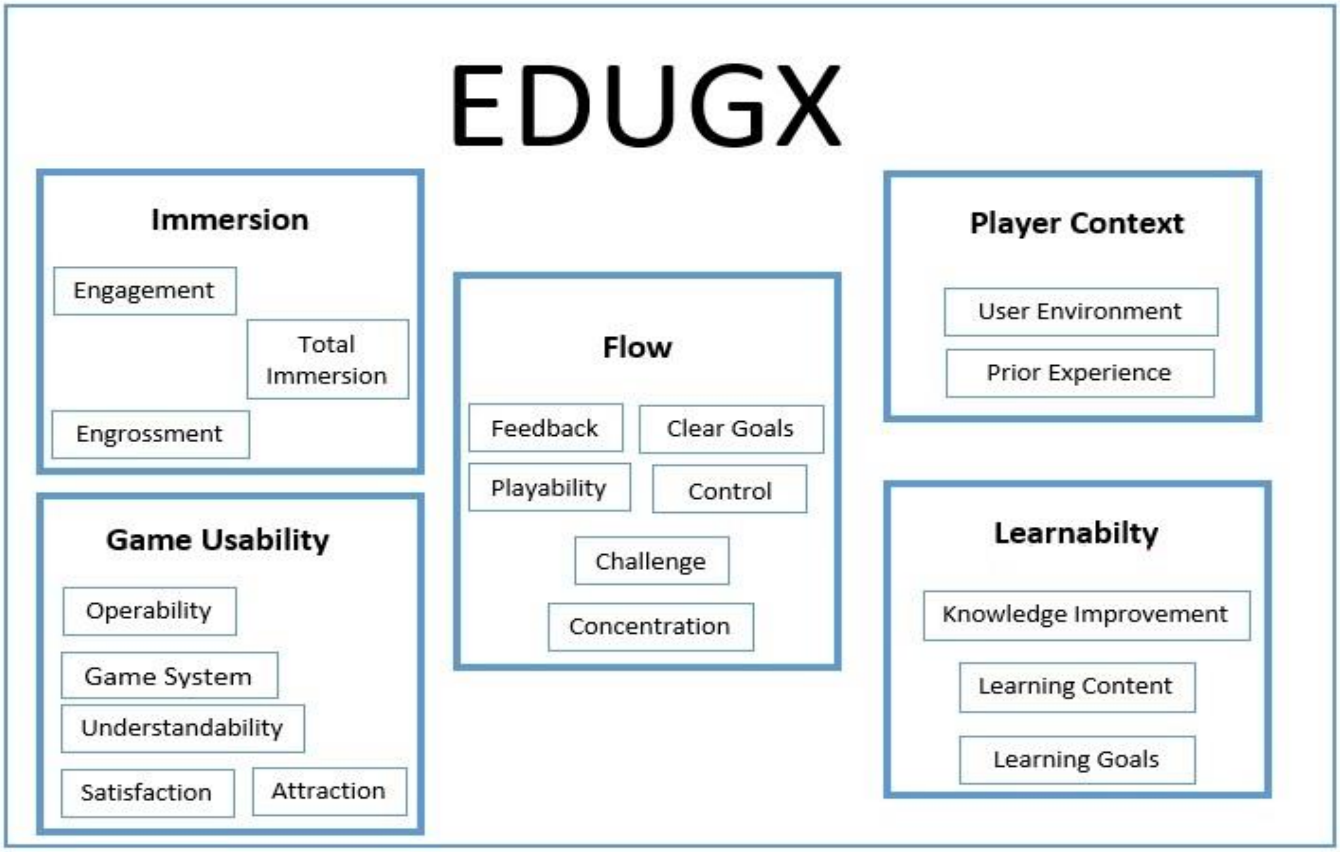


Fig. 1. Proposed EDUGX Framework.

As for the EDUGX framework, based on the suggestions from the experts, the game system element is combined under game usability thus the number of elements are reduced from six to five. The experts (R1, R4 and R5) suggested that it will be better to have fewer elements to evaluate. Therefore, based on the review and analysis, a proposed EDUGX framework is illustrated in Figure 1.

V. Conclusions and Future Works

This study was carried out as it is applicable to the present as well as future requirements of suitable educational game industry and contributes the knowledge to user experience evaluation field. As mentioned in this study, user experience plays an important role to make the products of the games more efficient and reliable. By depending on the studies in the literature review and experts' review, an UX evaluation framework of the educational game (EDUGX) which consist of five main elements Flow, Immersion, Player Context, Usability and Learnability was developed. A suitable User Experience (UX) tool for an educational game will ease the game designers work as well as contribute to effective educational games. Thus, through this paper, some reviews are done on previously define User Experience (UX) tool which was based on EDUGX framework elements to identify the most approachable tool to evaluate UX design for educational games (EDUGXQ). As for future work, EDUGXQ will be used in a real experiment to evaluate the UX in an educational game through a process of mixed method research design.

References

[1] Molla, R. Mary Meeker's 2017 internet trends report: All the slides, plus analysis. 2017; Available from: https://www.recode.net/2017/5/31/15693686/mary-meeker-kleiner-perkins-kpcb-slides-internet-trends-code-2017.

[2] Boyle, E., T.M. Connolly, and T. Hainey, The role of psychology in understanding the impact of computer games. Entertainment Computing, 2011. 2(2): p. 69-74.

[3] Bjørner, T. and C.B.S. Hansen, Designing an Educational Game: Design Principles from a Holistic Perspective. International Journal of Learning, 2010. 17(10): p. 279-290.

[4] Phan, M.H., J.R. Keebler, and B.S. Chaparro, The development and validation of the game user experience satisfaction scale (GUESS). Human factors, 2016. 58(8): p. 1217-1247.

[5] McFarlane, A., A. Sparrowhawk, and Y. Heald, Report on the educational use of games. 2002: TEEM (Teachers evaluating educational multimedia), Cambridge.

[6] Steiner, C., et al. Evaluation of serious games: a holistic approach. in ICERI2015: 8th annual International Conference of Education, Research and Innovation. 2015. IATED Academy.

[7] Fierley, R. and S. Engl. User experience methods and games: lessons learned. in Proceedings of the 24th BCS Interaction Specialist Group Conference. 2010. British Computer Society.

[8] Mahlke, S., I. Lemke, and M. Thüring, The diversity of non-instrumental qualities in human-technology interaction. Mmi-Interaktiv, 2007. 13.

[9] Nagalingam, V. and R. Ibrahim. Finding the Right Elements: User Experience Elements for Educational Games. in Proceedings of the 2017 International Conference on E-commerce, E-Business and E-Government. 2017. ACM.

[10] Law, E.L.-C., et al. Interplay between user experience evaluation and system development: state of the art. in NordiCHI 2012: International Workshop on the Interplay between User Experience (UX) Evaluation and System Development (I-UxSED 2012), Copenhagen, Denmark, 14-17 October 2012. 2012.

[11] Jong, M.S., J.H. Lee, and J. Shang, Educational use of computer games: where we are, and what's next, in Reshaping Learning. 2013, Springer. p. 299-320.

[12] Sweetser, P. and P. Wyeth, GameFlow: a model for evaluating player enjoyment in games. Computers in Entertainment (CIE), 2005. 3(3): p. 3-3.

[13] Arnab, S., et al., The development approach of a pedagogically-driven serious game to support Relationship and Sex Education (RSE) within a classroom setting. Computers & Education, 2013. 69: p. 15-30.

[14] Burnes, D., et al., Maximizing Effectiveness of Educational Games through Gaming Experience. From andrewd. ces. clemson. edu/courses/cpsc414/spring14/papers/group5. pdf [accessed January 2015], 2015.

[15] Sánchez, J.G., N.P. Zea, and F.L. Gutiérrez. Playability: How to identify the player experience in a video game. in IFIP Conference on Human-Computer Interaction. 2009. Springer.

[16] M.T. Cheng, H.C.S.L.A.A., Game immersion experience: its hierarchical structure and impact on game-based science learning. Journal of Computer Assisted Learning, 2015. 31: p. 232–253.

[17] Cairns, P., A. Cox, and A.I. Nordin, Immersion in digital games: review of gaming experience research. Handbook of digital games, 2014. 339.

[18] IJsselsteijn, W., et al. Characterising and measuring user experiences in digital games. in International conference on advances in computer entertainment technology. 2007.

[19] Hamari, J., et al., Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. Computers in Human Behavior, 2016. 54: p. 170-179.

[20] Fisk, A.D., et al., Designing for older adults: Principles and creative human factors approaches. 2009: CRC press.

[21] Bernhaupt, R., User experience evaluation in entertainment, in Evaluating User Experience in Games. 2010, Springer. p. 3-7.

[22] Bevan, N., J. Carter, and S. Harker. ISO 9241-11 revised: What have we learnt about usability since 1998? in International Conference on Human-Computer Interaction. 2015. Springer.

[23] Peixoto, D.C., R.F. Resende, and C.I.P. Pádua. Evaluating software engineering simulation games: The UGALCO framework. in Frontiers in Education Conference (FIE), 2014 IEEE. 2014. IEEE.

[24] Cheng, K. and P.A. Cairns. Behaviour, realism and immersion in games. in CHI'05 extended abstracts on Human factors in computing systems. 2005. ACM.

[25] Nacke, L., J. Schild, and J. Niesenhaus, Gameplay experience testing with playability and usability surveys–An experimental pilot study. Playability and Player Experience, 2010.

[26] Kiili, K., et al., The design principles for flow experience in educational games. Procedia Computer Science, 2012. 15: p. 78-91.

[27] Nacke, L., A. Drachen, and S. Göbel, Methods for evaluating gameplay experience in a serious gaming context. International Journal of Computer Science in Sport, 2010. 9(2): p. 1-12.

[28] Nacke, L. and A. Drachen. Towards a framework of player experience research. in Proceedings of the second international workshop on evaluating player experience in games at FDG. 2011.

[29] Engl, S. and L.E. Nacke, Contextual influences on mobile player experience–A game user experience model. Entertainment Computing, 2013. 4(1): p. 83-91.

[30] Moreno-Ger, P., et al., Educational game design for online education. Computers in Human Behavior, 2008. 24(6): p. 2530-2540.

[31] Ibrahim, A., et al., Educational Video Game Design Based on Educational Playability: A Comprehensive and Integrated Literature Review. 2012.

[32] Van Staalduinen, J.-P. and S. de Freitas, A game-based learning framework: Linking game design and learning. Learning to play: exploring the future of education with video games, 2011. 53: p. 29.

[33] Lucero, A., et al. The playful experiences (PLEX) framework as a guide for expert evaluation. in Proceedings of the 6th International Conference on Designing Pleasurable Products and Interfaces. 2013. ACM.

[34] Takatalo, J., et al., Presence, involvement, and flow in digital games, in Evaluating user experience in games. 2010, Springer. p. 23-46.

[35] Law, E.L.-C. and X. Sun, Evaluating user experience of adaptive digital educational games with Activity Theory. International Journal of Human-Computer Studies, 2012. 70(7): p. 478-497.

[36] Kiili, K., Digital game-based learning: Towards an experiential gaming model. The Internet and higher education, 2005. 8(1): p. 13-24.

[37] Kiili, K., et al., Flow framework for analyzing the quality of educational games. Entertainment Computing, 2014. 5(4): p. 367-377.

[38] Mashapa, J., et al. Managing user experience–managing change. in 14th International Conference on Human-Computer Interaction (INTERACT). 2013. Springer.

[39] Nacke, L.E., M.N. Grimshaw, and C.A. Lindley, More than a feeling: Measurement of sonic user experience and psychophysiology in a first-person shooter game. Interacting with Computers, 2010. 22(5): p. 336-343.

[40] Shoukry, L., C. Sturm, and G.H. Galal-Edeen, Pre-MEGa: a proposed framework for the design and evaluation of preschoolers' mobile educational games, in Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering. 2015, Springer. p. 385-390.

[41] Shiratuddin, N. and S.B. Zaibon. Designing user experience for mobile game-based learning. in User Science and Engineering (i-USEr), 2011 International Conference on. 2011. IEEE.

[42] Norwani, N.M., et al., Development of Teacher Leadership Guiding Principles in Preparing Teachers for the Future. International Journal of Academic Research in Business and Social Sciences, 2016. 6(12): p. 374-388.

[43] Singha, S., et al., Use of Content validity index for selection of occupational safety factors for workers in manufacturing industry. 2015.

[44] Maxwell, K.L. and V.H. Wright, Evaluating the Effectiveness of Two Teaching Strategies to Improve Nursing Students' Knowledge, Skills, and Attitudes About Quality Improvement and Patient Safety. Nursing education perspectives, 2016. 37(5): p. 291-292.

[45] Polit, D.F., C.T. Beck, and S.V. Owen, Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Research in nursing & health, 2007. 30(4): p. 459-467.

[46] Fu, F.-L., R.-C. Su, and S.-C. Yu, EGameFlow: A scale to measure learners' enjoyment of e-learning games. Computers & Education, 2009. 52(1): p. 101-112.

[47] Brockmyer, J.H., et al., The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. Journal of Experimental Social Psychology, 2009. 45(4): p. 624-634.

[48] Boberg, M., et al. PLEXQ: towards a playful experiences questionnaire. in Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play. 2015. ACM.

APPENDIX

APPENDIX A: ELEMENTS IN RELATED QUESTIONNAIRES INSTRUMENT FOR GAMES IN EDUCATION

Related Frameworks/ Models	**Instrument**	**Elements**
Flow Framework	GameFlow Questionnaire (GFQ) [46]	• Concentration • Goal Clarity • Feedback • Challenge • Autonomy • Immersion • Social Interaction • Knowledge Improvement
Objectives, Activity, Accomplishment and Affect (OA3) framework	Game Engagement Questionnaire [47]	• Absorption • Flow • Presence • Immersion
Presence-Involvement-Flow framework	Experiment Virtual Environment Questionnaires (EVEQ-GP) [34]	• Involvement • Presence • Flow
Gameplay Experience Model	Game Experience Questionnaire (GEQ) [29]	• Immersion • Presence • Flow • Absorption • Dissociation
UGALCO	UGALCO Questionnaire [23]	• Game Experience • Adaptivity • Learning Experience • Usability • Communicability
Playful Experiences (PLEX) framework	Playful Experiences Questionnaire (PLEXQ) [48]	• Captivation • Challenge • Competition • Completion • Control • Cruelty • Discovery • Exploration • Expression • Fellowship • Humour • Nurture • Relaxation • Sensation • Subversion • Suffering • Thrill
Game Immersion Experience	Game Immersion Questionnaire (GIQ) [16]	• Engagement • Engrossment • Total Immersion

APPENDIX B: CONTENT VALIDITY INDICES (CVI)

Element	Item	Number in Agreement	CVI	Element	Item	Number in Agreement	CVI
Flow	CG1	5	1.0	Immersion	*EG1*	*3*	*0.6*
	CG2	5	1.0		EG2	5	1.0
	CG3	*3*	*0.6*		EG3	4	0.8
	CG4	*3*	*0.6*		EG4	4	0.8
	CG5	4	0.8		EG5	4	0.8
	FB1	5	1.0		EG6	5	1.0
	FB2	5	1.0		ER1	4	0.8
	FB3	5	1.0		*ER2*	*3*	*0.6*
	FB4	*3*	*0.6*		ER3	4	0.8
	FB5	*3*	*0.6*		ER4	4	0.8
	PL1	5	1.0		TI1	4	0.8
	PL2	5	1.0		TI2	4	0.8
	PL3	*3*	*0.6*		TI3	4	0.8
	PL4	*2*	*0.4*		*TI4*	*3*	*0.6*
	PL5	*2*	*0.4*	Player Context	UE1	5	1.0
	PL6	*2*	*0.4*		*UE2*	*3*	*0.6*
	PL7	5	1.0		UE3	4	0.8
	PL8	4	0.8		UE4	4	0.8
	PL9	5	1.0		UE5	5	1.0
	CT1	5	1.0		UE6	5	1.0
	CT2	5	1.0		PE1	5	1.0
	CT3	5	1.0		PE2	5	1.0
	CT4	*3*	*0.6*		PE3	5	1.0
	CT5	5	1.0		PE4	5	1.0
	CT6	*3*	*0.6*	Game Usability	OP1	4	0.8
	CL1	5	1.0		OP2	4	0.8
	CL2	5	1.0		OP3	5	1.0
	CL3	4	0.8		OP4	4	0.8
	CL4	4	0.8		UD1	5	1.0
	CL5	4	0.8		UD2	5	1.0
	CL6	4	0.8		*UD3*	*3*	*0.6*
	CL7	4	0.8		UD4	5	1.0
	CN1	5	1.0		ST1	5	1.0
	CN2	5	1.0		ST2	5	1.0
	CN3	5	1.0		ST3	5	1.0
	CN4	*3*	*0.6*		ST4	4	0.8
	CN5	5	1.0		AT1	5	1.0
	CN6	*3*	*0.6*		AT2	5	1.0
	CN7	5	1.0		AT3	5	1.0
					AT4	4	0.8
Element	**Item**	**Number in Agreement**	**CVI**	**Element**	**Item**	**Number in Agreement**	**CVI**
Learnability	KI1	5	1.0	Game System	DV1	4	0.8
	KI2	5	1.0		DV2	4	0.8
	KI3	*3*	*0.6*		*DV3*	*2*	*0.4*
	KI4	5	1.0		DV4	4	0.8
	KI5	5	1.0		DV5	4	0.8
	LC1	4	0.8	Learnability	LG1	5	1.0
	LC2	5	1.0		LG2	5	1.0
	LC3	5	1.0		LG3	5	1.0
	LC4	5	1.0		LG4	4	0.8
	LC5	*3*	*0.6*		LG5	5	1.0

Stemming Text-based Web Page Classification using Machine Learning Algorithms: A Comparison

Ansari Razali[1], Salwani Mohd Daud [2]
Faezehsadat Shahidi[4]
Department of Informatics FTIR
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia

Nor Azan Mat Zin[3]
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor
Malaysia

***Abstract*—The research aim is to determine the effect of word-stemming in web pages classification using different machine learning classifiers, namely Naïve Bayes (NB), k-Nearest Neighbour (k-NN), Support Vector Machine (SVM) and Multilayer Perceptron (MP). Each classifiers' performance is evaluated in term of accuracy and processing time. This research uses BBC dataset that has five predefined categories. The result demonstrates that classifiers' performance is better without word stemming, whereby all classifiers show higher classification accuracy, with the highest accuracy produced by NB and SVM at 97% for F1 score, while NB takes shorter training time than SVM. With word stemming, the effect on training and classification time is negligible, except on Multilayer Perceptron in which word stemming has effectively reduced the training time.**

Keywords*—*Web page classification; stemming; machine learning; Naïve Bayes; k-NN; SVM; multilayer perceptron

I. Introduction

The fast-growing number of websites in the World Wide Web (WWW) necessitates efficient methodologies to locate information from millions of web pages. Internet has become a huge repository of information and thereby web page documents need to be categorized to facilitate the indexing, searching and web pages retrieval by the search engine [1]. An automation of web pages classification can be achieved by using machine learning. Supervised machine learning algorithms are used for problem that has label and predefined categories. Features that will be the input to the machine learning algorithms is gathered through web data mining; a process of extracting patterns from web pages data [2], which comprises web pages content, hyperlinks or user logs usage.

For this article, only the web pages contents, specifically texts are used as the features–images and audios are discarded. This allows web pages classification to be carried out similar to plain text document classification; words in the web pages are vectorized and become the features that train the classifiers. Pre-processing procedures such as stop word removal and word stemming are commonly conducted before running a machine-learning algorithm to reduce classifiers' processing time by reducing the features in the document. However, word stemmer is known to produce errors to the resulting stemmed words and this may affect classifiers' classification accuracy, which is measured by its precision and recall value [3].

The BBC dataset used in this research consists of news articles that are predefined and labeled based on five categories. Machine learning algorithms are used to extract and learn prominent features that defines each category so that future articles can be classified automatically. In webpages searches, the speed of classification process is an important factor that affects user experience. While this is important, faster processing should not be justified on the expense of classification accuracy. Thereby, this research evaluates word stemming procedure on the classification speed and accuracy, using different machine learning algorithms.

II. Web Page Classification

Web page classification, or also called web page categorization, is defined as a task to determine the category of a web page. In a formal definition, let C = {c1,…ck} as predefined categories, D = {d1,…dk} as web pages and A = D x C as a decision matrix (Table I).

whereby each entry a_{ij}, (1 <i<N, 1<j<K) indicates whether web page di is in category c_j. Each $a_{ij} \epsilon \{0,1\}$; 1 is when a web page di fits category c_j, 0 when it is not in c_j. A web page can be fitted in one category, multiple categories or none of the categories. The objective of web page classification is to estimate the unknown assignment function f: D x C → {0,1} by means of a learned function f': D x C → {0,1}, which is either a classifier, model or hypothesis, such that f' coincides with f to maximum extent. The learned function f' is derived from performing machine learning over a training data which consists of web pages that are labeled with their assigned categories. The trained function f' will then be used to classify unseen data of web pages to its categories [23].

TABLE. I. Decision Matrix

Web Pages	Categories				
	C_1	…	C_j	…	C_k
d_1	a_{11}	…	a_{1j}	…	a_{1k}
…	…	…	…	…	…
d_i	a_{i1}	…	a_{ij}	…	a_{ik}
…	…	…	…	…	…
d_n	a_{n1}	…	a_{nj}	…	a_{nk}

Web classification is almost similar to text classification, but with additional steps because of special characteristics in web pages:

- Web pages are semi-structured documents commonly written in HTML that has information enclosed between tags.
- Web pages have topological information about the link graph which shows hyperlinks information with the linked web pages.

In web page classification, there are multitude of potential inputs that can be used by classifier, such as URL of the web page, HTML tags frequency, the content of the tags and so on.

III. Related Works

Previous works on these algorithms use Naïve Bayes to classify 4,887 website homepage contents into 10 categories yielding 89% accuracy [4], k-NN shows higher accuracy as compared to Naïve Bayes for text and document classification despite showing low performance in terms of its fully dependency on every sample in the training set [5]-[6], SVM performed better than Naïve Bayes in classifying health and non-health related websites [3], Naïve Bayes trumps k-NN and SVM when classification is carried out to predict users' personality based on Twitter texts [5], and automatic web page categorization on educational based corpus is conducted using seven classifiers, with high accuracy classifiers demonstrated by Linear SVM, Logistic Regression, Multinomial Naïve Bayes, and Multilayer Perceptron. Decision Tree is the worst performed while k-NN is moderate.

An approach was used by extracting information from both web pages contents and links structure as inputs to SVM and neural network [7]. An improved k-NN classifier uses new feature weighting and new distance weighted voting scheme [8], and an improvement is suggested on the k-NN to adopt density-based approach to manage unevenly distributed dataset. The distance between k-NN and test data are adjusted based on their difference of density [9].

The effect of word stemming to the performance of text classification is arguable. A performed system should acquire high number of relevant documents (high recall) and only a few non-relevant documents (high precision). An evaluation of Porter stemming based on information retrieval from a corpus of 400 MEDLINE (Medical Literature, Analysis and Retrieval System Online) shows improvement of precision and recall as compared to information retrieval without using stemmer [10].

Researchers [11] argue that stemming has little impact on the performance of text classification. Schofield *et al* [12] have conducted experimental procedure to validate the outcome of various stemmers on different type of text corpus. The study concludes that generally stemmer yield no meaningful improvement in likelihood and coherence and can even degrade topic stability. The researcher claims that Porter stemmer for instance just reduce the possible unigrams that can be generated and does not appear to improve the model quality. Statistical approach of stemming does not need to have built in set of morphological rules as in rule-based approach; it learns the rule by training on a well-formed corpus. Thus, it overcomes Porter's error. Nonetheless, statistical approach has shortcomings such as dependency on corpus size and quality, higher execution time and high storage use [13]-[14]. In this report, we examine the effect of stemming to classifiers performance.

Previous works recorded extensive discussions on web pages classification using various types of machine learning algorithms. Many literatures however focus solely on the classification accuracy and does not include the processing time in the results. Additionally, even less literature records the difference in the accuracy and processing time with and without word stemming.

IV. Experiments

A. Dataset

This study uses dataset that originates from BBC News website articles gathered by [15]. It consists of 2,225 documents that corresponds to articles on five topical areas published on the BBC News website from year 2004 to 2005. The articles are labeled based on the topics, namely 'business', 'politics', 'entertainment', 'sport' and 'tech'. The BBC dataset consists of 2,225 documents and is split randomly into training and test dataset with the ratio of 80:20, which is a common ratio used for this purpose [16]. After splitting, there are 1,780 documents in training and 445 in test dataset. The frequency distribution for the dataset before and after splitting is shown by Table II.

This dataset comes in the form of raw text files and separated into five different folders based on their respective topics. Data in these text files need to be collated in a spreadsheet to enable further analysis and processes. Python codes are used to combine all the data and subsequently exported into a comma-separated values (.csv) spreadsheet file. A column named as 'newstype' is created to indicate the news category while column 'news' stores the news article.

B. Pre-Processing

The number of rows of a matrix corresponds to the number of words in a document collection. There can be hundreds of thousands of different words in the document collection. Pre-processing is an effort to reduce these words, which is the input feature for the machine learning classifiers, by cleaning up the document, selecting and extracting feature word and perform word stemming. These processes can improve computational efficiency and classification effectiveness.

1) Cleaning up document: Text documents are broken down into individual words through tokenization. Tokenization is commonly followed with other pre-processing steps such as removing stop words, punctuation, special characters and word stemming. These individual words will be selected and extracted to become features that represent respective categories or labels. These words or features serve as inputs to machine learning classifiers. After pre-processing, there are a total of 389,548 features extracted from the whole document collection. The process of cleaning text is important to remove unnecessary and non-important elements of sentences as show in Fig. 1.

TABLE. II. FREQUENCY DISTRIBUTION

News type	Number of documents (Before splitting)	Number of documents (After splitting)
Business	510	399
Entertainment	386	312
Politics	417	331
Sport	511	403
Tech	401	335

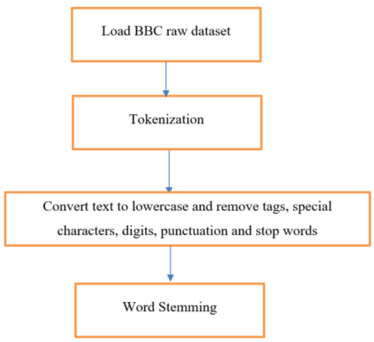


Fig. 1. Cleaning up Process of Text Document.

2) Feature extraction and selection: Feature selection and extraction is a process to reduce noise terms that are not related to the categories of training document. To reduce these noise terms, feature selection is first conducted to extract index terms (features) that will become the predictor to evaluate and assign unseen documents as belonging to the appropriate category. Since feature selection reduces noise terms, in effect it reduces vector dimensions and thus enable classifiers to produce faster results. Training documents of similar categories are represented with the same term vectors, thus they can be closely located in term vector space.

For the first step in text categorization, we need to transform the documents consisting of strings of characters into a representation that is suitable for the learning algorithms and the classification tasks. And the most commonly used document representation is Vector Space Model (VSM), that is, each document is represented by a vector of words. A word-by-document matrix A is used for a collection of documents, with each entry represents the occurrence of a word in a document, that is, A=(a_ij), where (a_ij) is the weight of word i in document j.

The weight value of each term can be computed by different weighted schemes namely Boolean value, Term Frequency (TF), Inverse Document Frequency (IDF), Term Frequency and Inverse Document Frequency (TFIDF) [17]. The simplest approach of determining the weight is Boolean weighting, which sets the weight (a_ij) to 1 if the word occurs in the document and 0 otherwise. TF weighted scheme counts the words that are most frequently occurring as shown by Fig. 2 which summarize the word count frequency from each of the five categories in the BBC dataset.

3) Word stemming: Stemming is a feature term reduction technique that is used by removing suffixes such as 'ed', 'ing' and 'ily'. It reduces complexity and enable more efficient information retrieval especially in data mining applications. Nonetheless, stemming may create non-real words as the stemmer does not check on grammatical rules during the stemming process [18]. Lemmatization is an alternative that checks on canonical forms of the words, but it is computationally expensive and thus takes up more processing time [19]. Porter stemmer is one of the most widely used stemmer. Other types of stemmer include Lovins, Lancaster and Porter2, which is also referred as Snowball [12]. Porter Stemmer algorithm as shown in Fig. 2 is commonly used in text classification. It is based on steps by which each step removes a type of suffix by using substitution rules. Non-real words such as 'studi' is a grouping stemmed words that resulted from words that comes from a similar root namely 'studied, studies, study, studying' [12],[20],[21].

Feature selection and extraction through applying TF-IDF and Porter Stemming are able to reduce dimensionality by trimming down the number of features from 389,548 to 233,123 features.

C. Classification

There are 1,780 documents in training and 445 in test dataset. The training dataset is used to train the machine learning classifiers whereby the classifier learns the characteristics of the dataset and the relations between these documents and their predefined categories. Once the classifier is fully trained, the test dataset is fed into the classifier as the input and it will output the predicted categories for each document. We compare few machine algorithms for the classification stage, i.e. Naïve Bayes, k-NN, Support Vector Machine (SVM), and Multilayer perceptron.

1) Naïve bayes: Naïve Bayes uses vector analysis method that is based on the concept of conditional probability or Bayes theorem to measure documents relevancy. The probability of class a to be the category for document b is given by:

$$P(a|b) = \frac{P(a)P(b|a)}{p(b)} \quad (1)$$

From the training dataset, the classifier calculates the probability value of each feature term belonging to certain category. It is based on the fraction of time a term appears among all terms in documents of a category. The sum of probabilities for each category of each term occurring in the document is calculated that enables the classification of a new document. This classifier is called 'naïve' because each term is assumed to occur independently from each other [6].

There are several types of Naïve Bayes, among them are Multinomial, Boolean and Bernoulli. For text classification, Multinomial Naïve Bayes is mostly used due to its computational efficiency and relatively good predictive performance. It uses multinomial distribution with the classification features consisting of the number of word occurrences or the weight of the word [22]. In this project, the weight of the word is used which is obtained from TF-IDF.

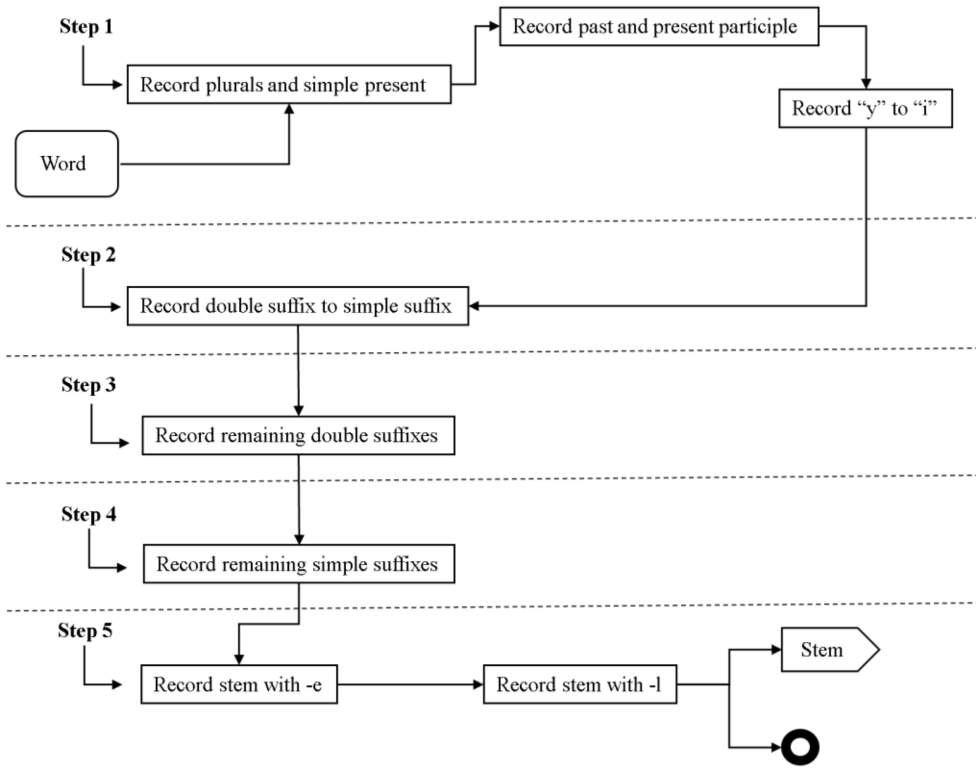


Fig. 2. Five Steps of Porter Stemmer [10].

2) k-Nearest Neighbor (k-NN): is a non-linear lazy learning classifier that delays learning process until a new document appear to be classified. It compares the new document directly with the training documents and computes their similarity score by measuring the distance between the documents by using Euclidean distance or cosine similarity. Euclidean distance is used in this project as it the most widely used distance metrics in k-NN classification. K-NN use the similarity score to rank the document's neighbors among the training document vectors. The k-nearest neighbors are used to predict the category of the new document [23].

3) Support Vector Machine (SVM): Support Vector Machine (SVM) is a powerful technique for classification. The state-of-the-art in text classification usually applies machine learning techniques such as SVM [24]. However, SVM is not suitable for large datasets or text corpora, because the training complexity of SVM is highly dependent on the input size [25]. Comparing the processing time, SVM takes longer time than NB and k-NN during classifier training but is faster than k-NN during classification. SVM is a universal learner. It is a linear learner in its basic form, but can be configured to learn polynomial classifiers, radial basic function (RBF) networks and three-layer sigmoid neural nets by applying appropriate kernel function. In a classification procedure carried out on big text corpora, [25] concluded that RBF and Sigmoid kernels need higher time to build model and requires additional parameters as compared to linear SVM. It is difficult to determine its parameterization with imbalanced data.

4) Multilayer Perceptron Multilayer Perceptron (MLP): also known as Artificial Neural Networks (ANN), is a multilayer, feed-forward neural network that contains nodes at the input layer, hidden layer and output layer. Having these multi-layers allows MLP to learn non-linear functions. Neurons in the hidden layer and output layers have biases that acts as weight. The purpose of learning is to assign the right weights to these edges that minimize the cost function. By entering vectors, these weights can determine the output vector.

MLP trains using backpropagation error method. Backpropagation error is a supervised learning method that computes the error at the output and used by gradient descent optimization to adjust the edge weights by calculating the gradient of the loss function. This adjustment is repeated iteratively, and iteration ends when the output error is below the established standard.

The problem with training MLP is to minimize error function E which is defined by the sum of square differences over all data in the training set. A simplified equation for error function E for an MLP with weights n is given as:

$$\min_{w \in \mathbb{R}^n} E(W) \quad (2)$$

with w∈R^n is column weight vector with components w_1, ω_2,… ω_n. There are various approaches to improve the efficiency of error minimization process, and one of the common methods used for text classification is quasi-Newton, which uses second order derivative related information [26].

Activation functions is an important element in ANN. Its purpose is to convert input signal of a node into non-linear property before channeling the signal to output signal, which then will be the input to the next layer in the stack. Non-linearity enables ANN in modelling complicated, high dimensional and not linearly separable big dataset. There are various types of activation functions, among mostly used are sigmoid, tanh and Rectified Linear Units (ReLU). Activation functions using sigmoid and tanh is less suitable for learning because its small derivatives can lead to vanishing gradient; when the neuron's activation saturates at either tail of 0 or 1, the gradient at these regions is almost zero. This will cause very slow or no learning during backpropagation as the weights are updated with small values. In this respect, ReLU function is less susceptible to this vanishing gradient issue because it has an identity derivative in the positive region [27]. Any negative elements are set to '0'; with no exponentials, multiplication nor division operations. Its gradient computation is simple and, in this way ReLU can speeds up neural networks training.

Although the number of hidden layers and nodes are an important determinant in ANN performance and processing time, there is no standard method on their selection. One of the method applicable is try and error approach [28]. Hidden layer size is arbitrarily selected, and the outcome is observed.

V. Results and Discussions

Table III summarize the results obtained from the experiments done for word stemming and without word stemming. Dataset is classified using Multinomial Naïve Bayes with the best classification outcome is achieved by using parameter Laplace smoothing and without word stemming; F1 score of 0.97, training time of 0.02 seconds and classification time of 0.78 seconds.

The best classification outcome for k-NN is achieved by using parameter k = 31 and without word stemming; F1 score of 0.96, training time of 0.007 seconds and classification time of 0.91 seconds.

Linear SVM without word stemming provides the best score with F1 of 0.97, training time of 11.37 seconds and classification time of 3.14 seconds. RBF and Sigmoid by far performed worse than linear SVM.

ANN classification is carried out using ReLU as an activation function. Generally, word stemming results in lower F1 score but reduces the training time. ANN with 3 layer and each layer containing 1,000 nodes shows the best F1 but long training time of 716.95 seconds. A more balanced 2-layer ANN with 50 nodes each takes only 15 seconds of training time and 0.73 seconds of classification time. This is taken as the best classification parameter and outcome for ANN. Another observation is ANN with three hidden layers performs no better than with one hidden layer. It does however introduce complexity and extends the training time.

Based on the Table IV and Fig. 3, in terms of classifiers' performance in classification, generally all of the classifiers perform at a high F1 score. The difference is marginally very low between the classifiers. Naïve Bayes and SVM each score 0.97 while k-NN and Multilayer Perceptron each get 0.96. All of the classifiers performed better with higher F1 score without word stemming. The effect of stemming on the training and classification time (as shown in Fig. 4) is negligible on all classifiers, except for Multilayer Perceptron. Stemming effectively reduced training time in Multilayer Perceptron modelling phase.

TABLE. III. Most Common Words in Each Category

Category	Politics	Tech	Business	Entertainment	Sport
	Party	People	Company	Film	Win
	Labour	Game	Firm	Award	Game
	Government	Mobile	Market	Star	Play
	Election	Technology	Rise	Music	Time
	People	Phone	Sale	Win	Player
	Blair	Service	Bank	Band	England
	Minister	User	Share	Actor	Match
	Tory	Firm	Economy	Director	Team
	Plan	Music	Price	Oscar	Final
	Brown	Software	Growth	Album	Club

TABLE. IV. Classifier's Best Performance Comparison

Classifier	Stemming	Precision	Recall	F1	Training time (seconds)	Classification time (seconds)
Naïve Bayes	No	0.97	0.97	0.97	0.02	0.78
k-NN	No	0.96	0.95	0.96	0.007	0.91
SVM	No	0.98	0.97	0.97	11.37	3.14
Multilayer Perceptron	No	0.97	0.96	0.96	18.17	0.73
Naïve Bayes	Yes	0.94	0.95	0.95	0.02	0.70
k-NN	Yes	0.92	0.92	0.92	0.007	0.98
SVM	Yes	0.95	0.95	0.95	10.23	2.3
Multilayer Perceptron	Yes	0.94	0.93	0.94	15.00	0.60

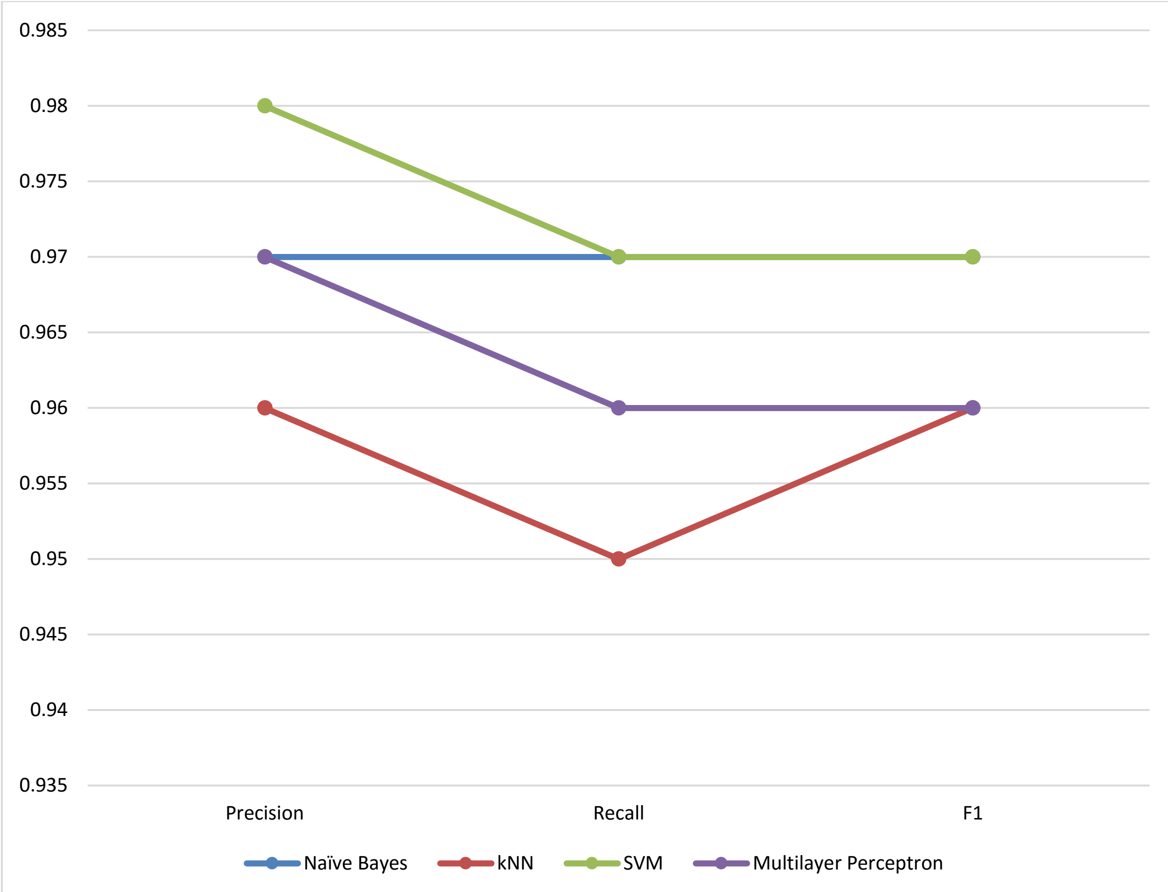


Fig. 3. Classifier Accuracy Comparison.

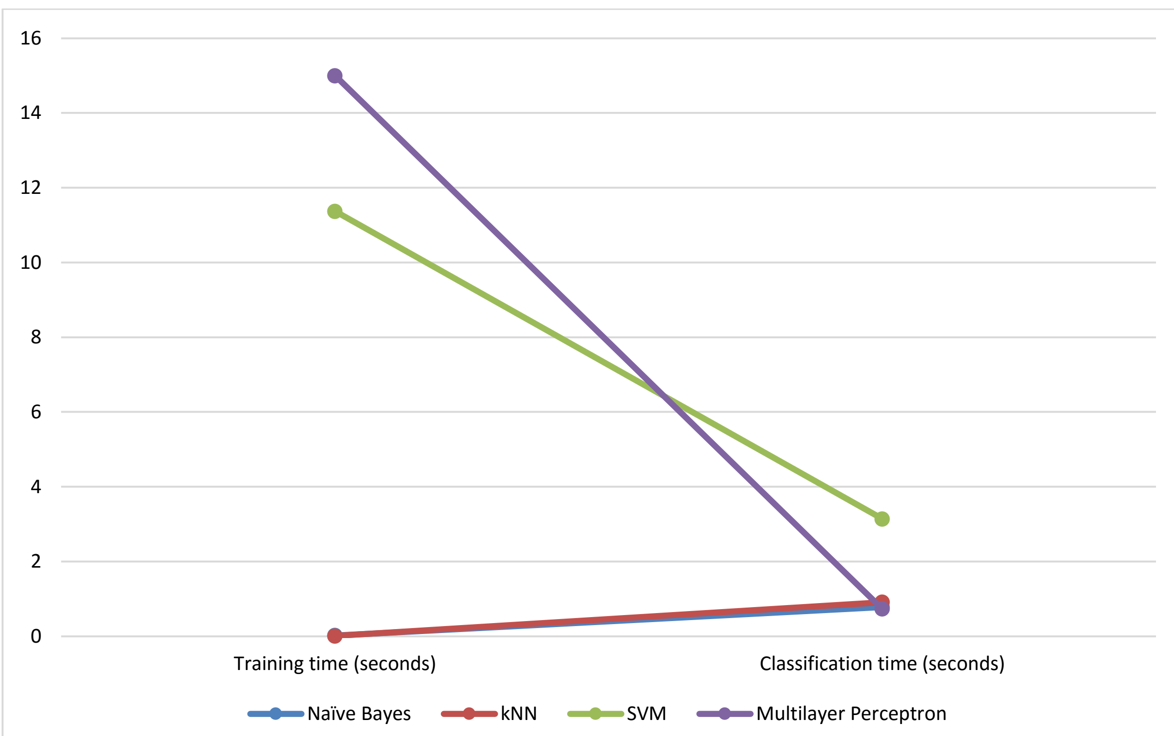


Fig. 4. Classifier Training and Classification Time Comparison.

VI. Conclusion

All of the classifiers produce respectable accuracy with very low marginal differences, which is not decisive. A larger dataset with larger number of categories may present higher complexity and dimensionality to the classifiers and perhaps with such challenges there will be a distinct best performer. This study exclusively uses Porter stemmer to perform word stemming. Future works may attempt the use of other word stemmer or lemmatization. Lemmatization may extend the processing time as it checks on canonical form of words. Nonetheless, it is interesting to observe on whether a correct grammar and real words produced from lemmatization will increase, instead of reducing, classifiers' accuracy as observed with Porter stemmer in this study.

Acknowledgment

The authors would like to express greatest appreciation to Ministry of Education, Malaysia and Universiti Teknologi Malaysia (UTM) for the financial support (grant number: Q.K130000.2538.19H12). Last but not least, we also appreciate Advanced Informatics at Razak Faculty of Technology and Informatics (FTIR), UTM for their support in conducting this research.

References

[1] A. Videira and N. Goncalves, "Automatic Web Page Classification Using Visual Content." in 10th Int. Conf. on Web Information Systems and Technologies, WEBIST 2014, Barcelona, Spain, April 2014, pp.279-294.

[2] M. Allahyari, S. Safaei, S. Pouriyeh, E.D. Trippe, K. Kochut, M. Assefi, and J.B. Gutierrez, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", J. ArXiv, vol. abs/1707.02919, August 2017.

[3] A. Siddiqui, M. Adnan, R. A. Siddiqui and T. Mubeen, "A comparative study of web pages classification methods applied to health consumer web pages," in 2nd Int. Con. on Computing Technology and Information Management, ICCTIM 2015, Johor, Malaysia, 2015, pp. 43-48.

[4] A.S. Patil and B.V. Pawar, "Automated Classification of Web Sites using Naive Bayesian Algorithm," in Int. MultiConference of Engineers and Computer Scientists, IMECS 2012, Hong Kong, March 2012, vol. 1, pp. 14-16.

[5] A. Moldagulova and R. Sulaiman, "Using KNN Algorithm for Classification of Textual Documents," in 8th Int. Conf. on Information Technology, ICIT 2017, Amman, Jordan, May 2017, pp. 665-671.

[6] V. Bijalwan, V. Kumar, P. Kumari, and J. Pascual, "KNN based Machine Learning Approach for Text and Document Mining", Int. J. of Database Theory and Application, vol. 7(1), pp. 61–70, 2014.

[7] M. Chau and H. Chen, "A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis," Decision Support Systems, vol. 44, pp. 482–494, 2008.

[8] J. Alamelu Mangai, Satej Milind Wagle, and V. Santhosh Kumar, "A Novel Web Page Classification Model using an Improved k Nearest Neighbor Algorithm" in 3rd Int. Conf. on Intelligent Computational Systems, ICICS 2013, Singapore, April 2013, pp. 50-53.

[9] K. Shi, L. Li, H. Liu, J. He, N. Zhang and W. Song, "An improved KNN text classification algorithm based on density," in IEEE Int. Conf. on Cloud Computing and Intelligence Systems, Beijing, 2011, pp. 113-117.

[10] W. Abdessalem, "A New Stemmer to Improve Information Retrieval," Int. J. Network Security and Its Applications (IJNSA), vol. 5, pp. 143-154, Jul 2013.

[11] M. Toman, R. Tesar, and K. Jezek, "Influence of Word Normalization on Text Classification," in InSciT, Oct 2006, pp. 354-358.

[12] A. Schofield and D. Mimno, "Comparing Apples to Apple: The Effects of Stemmers on Topic Models," Transactions of the Association for Computational Linguistics, vol. 4, pp. 287-300, 2016.

[13] X. Jinxi, W.C. Bruce, "Corpus-based Stemming Using Co-occurrence of Word Variants," ACM Transactions on Information Systems, vol. 16(1), pp. 61-81, 1998.

[14] D. Patel, M. Patel, and Y. Dangar, "A Survey of Different Stemming Algorithm," Int. J. of Advance Engineering and Research Development, vol. 2(6), pp. 50-53, 2015.

[15] D. Greene and P. Cunningham, "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering," in 23rd Int. Conf. on Machine Learning, ACM, Jun 2006, pp. 377-384.

[16] E. Setiani and W. Ce, "Text Classification Services Using Naïve Bayes for Bahasa Indonesia," in Int. Conf. on Information Management and Technology, ICIMTech, Jakarta, Indonesia, 2018, pp. 361-366.

[17] Singh, K. Nareshkumar, Devi, H. Mamata, Mahanta, A. Kakoti, "Document Representation Techniques and Their Effect on Document Clustering and Classification: A Review," Int. J. Advanced Research in Computer Science, vol. 8(5), pp. 1781-1784, 2017.

[18] Q.S. Mahdi, K.H. Qadir, and R.L. Falih, "Web Page Classification by Using Neural Networks," Zanco J. Pure Applied Sciences, vol. 23, 2011.

[19] S. Raschka, "Naïve Bayes and Text Classification I – Introduction and Theory." J. ArXiv, 2014.

[20] C. Naik and V. Kothari, "Document Classification using Neural Networks Based on Words," Int. J. Advanced Research in Computer Science, vol. 6(2), pp. 183-188, 2015.

[21] B.L. Devi and A. Sankar, "Feature Selection for Web Page Classification Using Swarm Optimization," Int. J. of Computer and Information Engineering, vol. 9(1), pp. 340-346, 2015.

[22] B.Y. Pratama and R. Sarno, "Personality Classification Based on Twitter Text Using Naïve Bayes, KNN and SVM," in Int. Conf. on Data and Software Engineering, Yogyakarta, Indonesia, 2015, pp. 170-174.

[23] B. Choi and Z. Yao, "Web Page Classification," Foundations and Advances in Data Mining, Springer, 2005.

[24] J.H. Wang and H.Y. Wang, "Incremental Neural Network Construction for Text Classification," in Int. Symposium on Computer, Consumer and Control, Taichung, Taiwan, 2014, pp. 970-973.

[25] R. Romero, E. L Iglesias, and L. Borrajo, "A Linear-RBF Multikernel SVM to Classify Big Text Corpora," BioMed Research Int. 2015.

[26] I.E. Livieris, "Improving the Classificaiton Efficiency of an ANN Utilizing a New Training Methodology," Informatics, vol. 6, p.1, 2019.

[27] S. Eger, P. Youssef, I. Gurevych, "Is it Time to Swish? Comparing Deep Learning Activation Functions Across NLP Tasks," in Con. on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018, pp. 4415-4424.

[28] F.S. Panchal, and M. Panchal, "Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network," Int. J. Computer Science and Mobile Computing, IJCSMC, November 2014, vol. 3, pp. 455 – 464.

Performance Realization of CORDIC based GMSK System with FPGA Prototyping

Renuka Kajur[1]
Assistant Professor, Department of Electronics and Communication Engineering, PESIT, Bengaluru, India

K V Prasad[2]
Professor and Head, Department of Electronics and Communication Engineering, BIT, Bengaluru, India

Abstract*—The Gaussian Minimum Shift Keying (GMSK) modulation is a digital modulation scheme using frequency shift keying with no phase discontinuities, and it provides higher spectral efficiency in radio communication systems. In this article, the cost-effective hardware architecture of the GMSK system is designed using pipelined CORDIC and optimized CORDIC models. The GMSK systems mainly consist of the NRZ encoder, Integrator, Gaussian filter followed by FM Modulator using CORDIC models and Digital Frequency Synthesizer (DFS) for IQ Modulation in transmitter section along with channel, the receiver section has FM demodulator, followed by Differentiator and NRZ decoder. TheCORDIC algorithms play a crucial role in GMSK systems for IQ generation and improve the system performance on a single chip. Both the pipelined CORDIC and optimized CORDIC models are designed for 6-stages. The optimized CORDIC model is designed using quadrature mapping method along with pipeline structure. The GMSK systems are implemented on Artix-7 FPGA with FPGA prototyping. The Performance analysis is represented in terms of hardware constraints like area, time and power. These results show that the optimized CORDIC based GMSK system is a better option than the pipelined CORDIC based GMSK systems for real-time scenarios.

***Keywords*—GMSK; CORDIC algorithm; FPGA; DFS; Gaussian Filter; pipelined; integrator; differentiator; channel**

I. Introduction

In the current era of high-speed communication, the prime objective of the system is to achieve modulation that has a power spectrum with a constant amplitude and adequate bandwidth. Out of all, some of the efficient techniques are Minimum Shift Keying (MSK) and Gaussian Minimum Shift Keying (GMSK). They are derived from Constant Phase-FSK family of modulation operating at a constant envelope. As a modulated signal has the characteristic of constant amplitude, the power consumption can be minimized by using a Class C RF amplifier. Employing an amplifier is necessary for battery operated units. Before modulation, the MSK signal along the I-Q component allows via half-sine shaper to pass the modulation signal. GMSK incorporates the same technique as that of MSK. The only variation is, instead of the shape of the half-sine pulse the inputs bits form the shape of the Gaussian bell curve. The realization of such shapers is carried using different digital or analog circuits of PCM LUTs [1-7].The present wireless digital communication systems use different types of modulation techniques to achieve high performance with efficient usage of available spectrum efficiency. The MSK provides the constant envelope signals for power amplification which reduces the problems caused by non-linear distortion, and MSK encodes each bit is like a half sinusoidal. The spectrum is not close enough to aware of the data rate approaches in RF channel Bandwidth. The GMSK overcomes the problems of MSK. GMSK is used in many wireless applications, and it limits the spectral bandwidth by using Gaussian filtering [1-2].

The GMSK is widely used in mobile communications and associated modulation scheme for GSM standards, and most of the GSM Mobiles have long battery life with greater efficiency and because of RF-Power amplifiers. The software-defined radio is a simplified approach which is adopted in most of the cellular standards. The American /European countries use GMSK modulation schemes in cellular phones as per GSM standards. The GMSK supports non-linear modulation, which is used for easy implementation on hardware [3-5]. If the Radio signal frequency increases in mobile communication and it is difficult to use by conventional technology like DSP, Microcontroller and Microprocessor in the real-time conditions due to its computational complexities. The GMSK system based transmitter and receiver is designed with timing synchronization using the Mueller and Muller method. The timing synchronization is necessary for real-time usage for GMSK systems, GMSK receiver consumes larger chip area and affects the overall system performance [6]. The GMSK modulation for EDGE mobile communication is established with a new digital technique of linear approximation based on Inter-symbol-interference (ISI) and partial response signaling. The GSM/Edge systems support narrow bandwidth, reasonable power spectrum usage, lesser the impulse response overshoot, higher immunity to noise interface and constant filter output pulse with a modulation index of 0.5[7]. The GMSK system is mainly used in Automatic identification systems (AIS) [8], OFDM communication systems [9] and digital image processing [10] and other applications.

Section II discusses the existing approaches of the GMSK system and its application usage and also the review of the new CORDIC algorithms and problem identification with research gaps. Section III describes the CORDIC algorithm principles along with pipelined and optimized CORDIC hardware architecture. The GMSK system is elaborated in section IV with GMSK principles, GMSK modulation, and Demodulation architectures. The results and performance analysis of Pipelined CORDIC and optimized CORDIC based GMSK systems are represented in section V. section VI concludes the overall work of GMSK systems with constraints improvements.

II. Existing Works

In this section, the existing approaches towards GMSK systems and its application usage and also a few existing CORDIC based approaches are elaborated. Munir et al. [11] present Cube Sat AIS receiver module using GMSK Modulation, the received AIS signals from the transmitter are at low power, and it is challenging to analyze operating frequency. So by deploying the GMSK modulation, the signals are analyzed inappropriate conditions. Poletaev et al. [12] present phase distortion variation estimation of GMSK modulated signals under Very low frequency (VHF). The phase distortion detector is designed using VCO's for IQ modulation and arc tan module for IQ generations for the generation of phase drift of VLF signals. Rakesh et al. [13] present the software-based approaches for BER analysis of GMSK system under AWGN channel condition. The Simulink tool is used for GMSK systems which include an encoder, GMSK Modulator followed by AWGN channel and GMSK Demodulator along with Viterbi decoder. The GMSK signals are generated for mobile radio telephony using Simulink modeling by George et al. [14]. The performance analysis of the GMSK Model for different fading channel conditions along with hard decision and soft decision decoding. The soft decision decoding techniques give better BER than hard decision decoding. Ghnimi et al. [15] present GMSK modulation under radio mobile propagation conditions based on the Matlab environment and its performance realization including AWGN along with one path and four path channel fading are analyzed.

The FPGA Based GMSK Modulator is designed by Nitin et al. [16] for GSM system which includes differential encoder-decoder, ROM based sine and cosine wave (IQ) generation using phase trajectory, phase concatenation followed by phase accumulation module. Gupta et al. [17] present GMSK transceiver on the FPGA platform for VHF waveform generation. The transmitter design includes Gaussian filter followed by FM Modulator with Direct digital synthesizer (DDS) up converter using XILINX system generator tool. Similarly, for GMSK receiver, DDS followed by Cascaded-integrator-comb (CIC) decimator filter and non-coherent Viterbi decoder. The GMSK carrier phase recovery loop for non-linear analysis are evaluated by jhaidri et al. [18], and the system includes precoder, GMSK Modulator followed by AWGN channel, along with coherent demodulation and carrier phase recovery on Matlab simulation environment. The FPGA based GMSK Demodulator using CORDIC engine is designed by Kumar et al. [19], and demodulator includes IQ generation using 1024 x16 ROM Module followed by 6-stage CORDIC module along with differentiator and decision synchronizer module. The design consumes a huge amount of chip area and power.

The floating-point Arthematic processor based on CORDIC algorithm is designed by bingyi et al. [20] which includes mantissa bits generation by pre-processing unit, followed by reconfigurable CORDIC rotation unit and exponent bits for normalization using post-processing model. The CORDIC rotation unit is a unified structure for multiplication, division, and square roots. The Hoang et al. [21] presents 32-bit Floating point FFT twiddle factor calculation unit using adaptive CORDIC module. The CORDIC module generates the sine and cosine values are used for FFT twiddle factor calculation for each iteration. Madi et al. [22] present a hardware implementation of sine and cosine generation using CORDICmodule. The phase inputs are generated from the ROM table followed by CORDIC5.0 is adopted form Xilinx IP Coreis modeled using Xilinx system generator. The Mousetrap-radix-2 CORDIC module is designed by chagela et al. [23], which includes asynchronous pipelined architecture to improve the throughput and power-delay product.

Research Gap:It has been noticed from the review of existing works carried on GMSK systems are based on software and few on hardware-based approaches. The recent existing approaches towards GMSK systems are not optimized yet and facing computational complexity, performance degradation, more resource utilization on hardware. The existing GMSK Modulation uses Voltage controlled oscillator (VCO), or local oscillators for IQ modulation or ROM based LUT, or current waveform generation approaches for IQ generation, which affects the overall chip area and system performance on GMSK systems. Few approaches are used CORDIC based designs for IQ generation but are facing problems on angle rotation convergence, more iteration stages to achieve the sine and cosine values, not pipelined and LUT based memory for tangent values updating. These research gaps are overcome by using the Proposed CORDIC based GMSK system in next sections.

III. CORDIC Designs

The CORDIC algorithm is an algorithm to calculate the trigonometric and hyperbolic operations, and it is also known as volder's algorithm. CORDIC typically intersect with one digit per iteration. In this section CORDIC algorithm principle and hardware architecture of pipelined and optimized CORDIC is explained in detail.

A. CORDIC Algorithm Principles

The system receives a vector and rotates by an angle θ for each iteration is

$$\begin{aligned} x_A &= x_i \cos\theta - y_i \sin\theta \\ y_A &= x_i \sin\theta + y_i \cos\theta \end{aligned} \tag{1}$$

If we consider, $x_i = 1$ and $y_i = 0$, after the first rotation, it will be

$$\begin{aligned} x_A &= \cos\theta \\ y_A &= \sin\theta \end{aligned} \tag{2}$$

To calculate the sine and cosine angles through rotation mode. Simplify the equation (1) to

$$\begin{bmatrix} x_A \\ y_A \end{bmatrix} = \cos\theta \begin{bmatrix} 1 & -\tan\theta \\ \tan\theta & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \tag{3}$$

One rotation is performed using the equation (3), and it requires four multiplications along with addition and subtractions. To avoid the multiplications, use small arbitrary angles by $\tan\theta_i = 2^{-i}$ for $i = 0,1,,,n$. Simplify the equation (3) to i[th] iterations is

$$x_{i+1} = x_i - S_i 2^{-i} y_i$$
$$y_{i+1} = y_i + S_i 2^{-i} x_i$$
$$z_{i+1} = z_i - S_i \tan^{-1}(2^{-i}) \tag{4}$$

The angle of the final iteration values accumulated by equation (4) and compared with an initial value z_0 . The sign S_i belongs to +1 for counterclockwise direction and -1 for clockwise direction.

B. Pipelined CORDIC Model

The CORDIC model is an iterative structure. So it will take a longer time to process an algorithm for the number of iterations. So use the pipelined registers between the iteration stages to achieve the fast computations. The sine and cosine angles have been computed in a row using a pipelined structure. The hardware architecture of the pipelined CORDIC model is represented in figure 1. For the rotation mode, set the initial values to x_0 = 1 and y_0=0 and z_0= θ. The pipelined structure has a 6-stage (i=6) iterative process. The 8-bit CORDIC model has pipelined registers, shifters, adder/subtract or along with constant tangent values with counter mode. The accuracy of the computation will be increased based on the number of iterative stages. These initial values are shifted by i-bits, where i is an integer and shifted up to 0 to 5. The division of X_i and Y_i by 2^i is taken place by the shifter for each stage. The new vector values are generated at intermediate stages for the given vector, and it is iteratively rotated to get the desired angle Z_i. The Z_i will decide the sign for addition and subtraction. The selection of the sign $S_i$set by

$$S_i = \begin{cases} -1 if\ Z_i < 0 \\ 1\ if\ Z_i > 0 \end{cases} \tag{5}$$

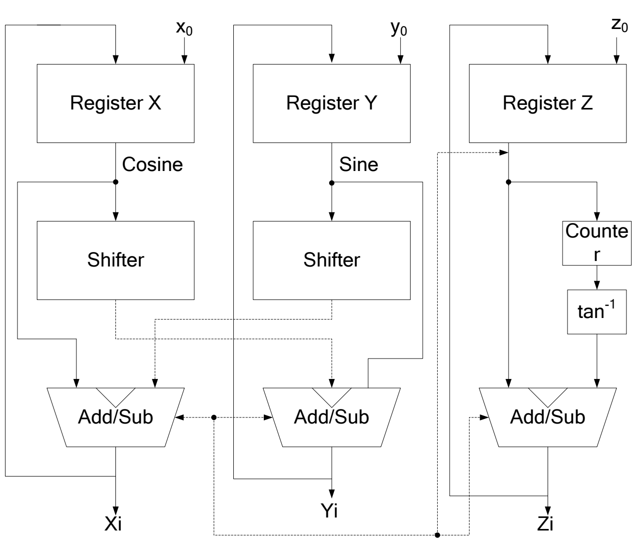


Fig. 1. Hardware Architecture of the Pipelined CORDIC Model.

If the initial stage is set to x_0 = 1 and y_0=0, then results are in the form of discrete sine and cosine values. It is difficult to realize on FPGA. So varies the Xi and Yi values to realize on hardware and to avoid the fractional values. In order to improve the accuracy, multiply the constant K= 0.611 with 2^6, so the initial values will set toX_i = 38 + x_0 and Y_i=0+ y_0 for six iterations. The addition/ subtraction outputs are feedback to pipelined registers to get each stage CORDIC outputs of sine and cosine values.

The tangent values are generated based on the following equation.

$$ZT_i = 2^6 \tan^{-1}(2^i) \tag{6}$$

The tangent values (ZT_i) are assigned for successive iteration using the counter to generate the $Z_i$values. The counter will count till 5 and reset to 0.

C. Optimized CORDIC Model

The improvised version of the pipelined CORDIC model is designed in this section and calls it as optimized CORDIC model. The significant difference between the above pipelined and this optimized CORDIC is quadrant mapping for proper angle rotation using preprocessing and post-processing, and separate pipelined structure for sine and cosine calculation and also delay module is used to synchronize the quadrant mapping with pipeline structure. The optimized CORDIC model is designed and represented in figure 2. The pipelined structure has a part of optimized CORDIC model is represented in figure 3.

The optimized CORDIC is overcome the rotation angle range problem, computational complexity, and avoiding the ROM table usage for tangent calculation, which saves the chip area of the CORDIC model. This optimized CORDIC is one of the best solutions for sine and cosine signal (IQ) generation to improve the speed and accuracy for GMSK systems and its applications.

The optimized CORDIC model has three stages namely, preprocessing, pipelined structure, and post-processing. The preprocessing stage receives the 8-bit phase input and for quadrant mapping, chooses the MSB 2-bit phase input [7:6] and transformed to the first quadrant. The same 2-bit is passed to the delay unit. The MSB of phase 2-bits [7:6] is "00", then assign the 8-bit phase input (PI) into phase register as a first quadrant. Similarly, if the bits are "01", assign PI-90^0, for "10", assign PI-180^0 and for "11", assign PI-270^0 into phase register as a first quadrant. This 8-bit phase register is input to the pipelined CORDIC structure.

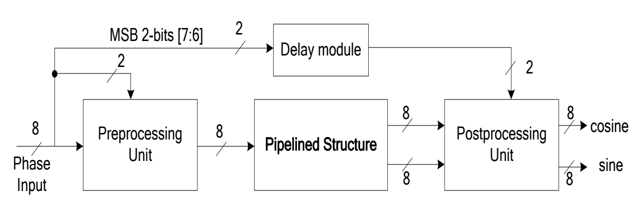


Fig. 2. Optimized CORDIC Model.

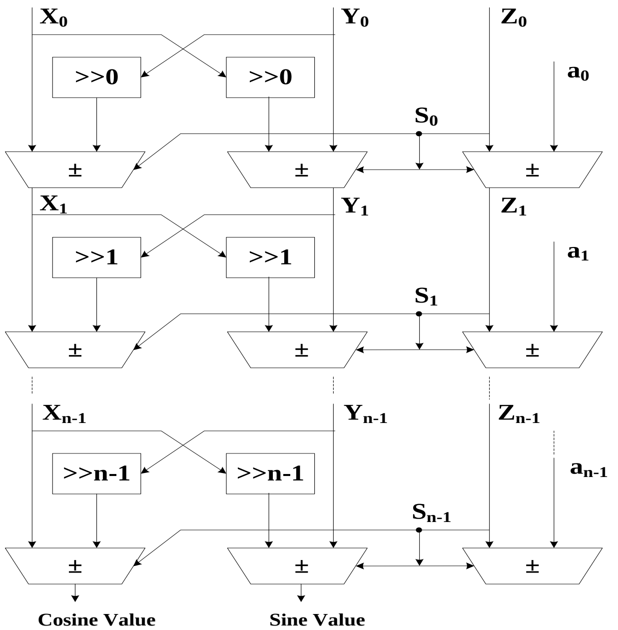


Fig. 3. Pipelined Structure used in Optimized CORDIC Model..

The pipelined structure is proportional to the accuracy of the angles and provides the high-speed calculation. In pipelined structure, Initialize the X_0= 38 (Decimal value), $Y_0 = 0$ and Z_0 = 8-bit phase register value as angle input before the iteration starts. The pipelined structure has six stages, and set n=6 is shown in figure 3. The phase register input Zi [7] decides the sign(S) for the corresponding iterations stages. The constant tangent values (a_0 to a_5) are assigned directly while performing the iterations. The adder/subtract or of each level completes the corresponding iteration stages. The pipelined registers are placed (not shown in figure 3) after each addition/subtraction operation except the final iteration stage. In six stage, the sine and cosine values are generated.

The delay register receives the MSB 2-bit phase input and generates the delayed 2-bit quadrant bits results parallelly with pipelined structure outputs to maintain the synchronization problems. The post-processing unit converts the first quadrant of the phase input to next quadrants along with pipelined outputs. The post-processing quadrant 2 bit generates the final cosine and a sine value from 6th stage pipelined outputs is as shown in table 1.

TABLE. I. QUADRANT MAPPING

Quadrant	Angle range	Sine value	Cosine value
00	$(0, 90^0)$	sinθ	cosθ
01	$(90^0, 180^0)$	cosθ	~sinθ
10	$(180^0, 270^0)$	~sinθ	~cosθ
11	$(270^0, 360^0)$	~cosθ	sinθ

IV. PROPOSED GMSK SYSTEM

In this section, the proposed GMSK system methodology is addressed using the pipelined CORDIC and optimized CORDIC model and its design flow is represented in figure 4. The proposed work main aims to design cost-effective hardware architecture of the GMSK system using pipelined CORDIC and optimized CORDIC models on a single chip. The CORDIC models are incorporated for IQ generation, which speeds up the GMSK systems in real time scenarios. The GMSK Systems mainly consists of Non-return to- Zero (NRZ) encoder-decoder, Integrator- Differentiator, Gaussian filter, FM Modulator, and FM Demodulator along with Channel. The GMSK design has 1-bit input data feed serially one after other, and process the GMSK operation, generates the 1-bit output, which is similar to GMSK input data.

A. Algorithm Principle

The mathematical representation of the GMSK modulated signal is explained in this section.

1) Consider the binary data stream $a(n), a \in (0,1)$ and convert to an antipodal sequence $b(n), b \in (0,1)$ and a stream of rectangular pulses *r(t)* is

$$r(t) = \sum_n b(n) p(t - nTb) \tag{7}$$

Where pulse $p(t) = 1, t \in (0, Tb)$, $p(t) = 0, otherwise$ $Tb = 1/fb$ and is symbol interval.

2) The integrator is single-pole infinite impulse response with unity feedback coefficient, and in general, it is represented as

$$i(n) = i(n-1) + r(t) \tag{8}$$

Where n =0, 1, 2, etc., many delay elements used in the integrator.

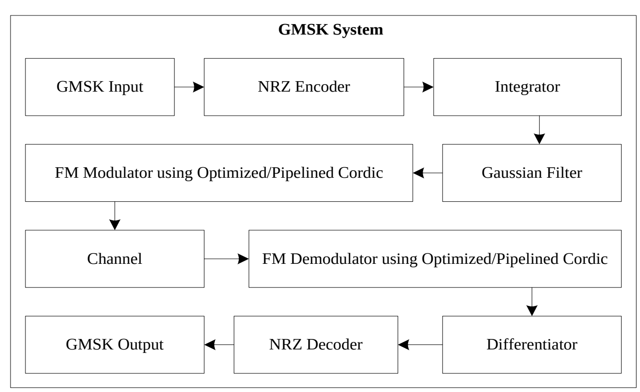


Fig. 4. Methodology of the Proposed Work.

3) The Gaussian filtering is one of the filtering technique which minimizes the group delay, and Gaussian function is used as an impulse response of the Gaussian filter. The Gaussian filter is designed using the FIR method and is expressed as

$$g(t) = \sum_{t=0}^{M-1} h(t).i(t) \tag{9}$$

Where i (t) is filtered pulse, M=8 for tap filter and Gaussian filter is specified by BT product, and impulse response h (t) is

$$h(t) = \frac{1}{\sqrt{2\pi}.\delta T} e^{\frac{-t^2}{2\delta^2 T^2}} \tag{10}$$

Where standard deviation $\delta = \frac{\sqrt{\ln 2}}{2\pi BT}$ and BT =0.3 for GSM

4) The modulation of the carrier wave frequency by $g(t)$ around a center frequency f_c implies modulating its phase by integrating the filtered output. so

$$\phi(t) = W_C t + 2\pi f_m + \int_0^t g(t).dt \tag{11}$$

Where the phase has been normalized to '0' at $t = 0$. The instantaneous frequency of the modulating signal will be

$$f(t) = \frac{1}{2}\frac{d\varphi}{dt} = f_c + f_m g(t) \tag{12}$$

Where f_c is carrier frequency and f_m is peak frequency deviation.

5) The peak frequency deviation is measured by the bit rate, and the GMSK need a one-bit interval of the period T_b contains '*N*' cycles of frequency $f_c - f_m$ and $N + \frac{1}{2}$ cycles of $f_c + f_m$. This leads to $f_m = \frac{1}{4T_b} = \frac{f_b}{4}$.

6) The modulated carrier is generated using the DFS method, and it can produce

$$y(t) = \cos(\phi(t)) \tag{13}$$

7) For maintaining the adequate sampling rates with digital methods requires at higher operating frequencies, So use Quadrature implementation approach to generate the modulated signal. The GMSK modulated signal $y(t)$ is

$$y(t) = \cos(W_C t)\cos(2\pi f_m \int g[T].dT) - \sin(W_C t)\sin(2\pi f_m \int g[T].dT) \tag{14}$$

Where $W_C = 2\pi f_c$ and Simplified version of the modulated signal $y(t)$ is

$$y(t) = I(t)\cos(2W_C t) - Q(t)\sin(W_C T) \tag{15}$$

B. Hardware Implementation

The GMSK Modulator mainly consists of NRZ Encoder, Integrator Gaussian filter followed by FM Modulation using CORDIC model's for IQ generation and DFS for IQ modulation and the hardware architecture of the GMSK Modulator is represented in figure 5. The GMSK modulator process 1-bit GMSK input serially and generates 12-bit GMSK modulated output.

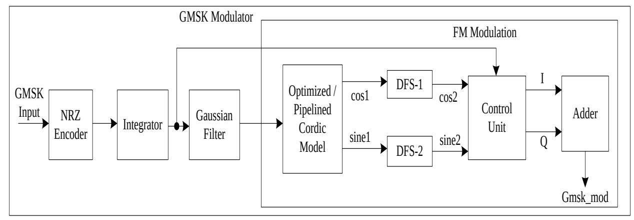


Fig. 5. Hardware Architecture of GMSK Modulator.

The NRZ Encoder receives the 1-bit GMSK input and generates the 1-bit output according to the NRZ logic. The NRZ encoder mainly contains the XOR module and data-Flip-flop. The XOR Module receives the GMSK input along with D-FF feedback output to process the transactions. The NRZ encoder is mainly used for slow speed transmission which interfaces for the synchronous and asynchronous process. The NRZ is to Logic '1', means the bit is set as a high and Logic '0', means the bit is set as a low value. The NRZ encoded output will be inverted and is input to Integrator, and It is mainly used for high data rate conversion with flexible multiplier less filter. The Integrator is used in the design is single-pole infinite impulse response (IIR) with unity feedback value. The integrator uses four data registers, and the NRZ encoded data is input to the first register. The first and fourth register outputs are added to generate the 1-bit integrated output.

The Gaussian filter is used to minimize the group delay in the GMSK process, and Gaussian function is the impulse response of the Gaussian filter. To design a Gaussian filter, the Low power FIR filter is used. The FIR design is an 8-tap filter, which is having 8- Gaussian filter coefficients, is multiplier with integrated output and generates the 8 temporary output values. These output values are stored in the eight registers parallelly. Using adder, add the corresponding register outputs individually, and 7th adder output is Gaussian filter output.

The FM Modulation mainly consists of pipelined CORDIC or optimized CORDIC model for IQ generation, two –Digital frequency synthesizer (DFS) modules for IQ Modulation, Control unit to generates I, Q separately, and an adder unit.

The pipelined CORDIC or optimized CORDIC models are explained in section III. The pipelined CORDIC or optimized CORDIC models receive the phase (angle) input and generate the In-phase (cos1) and Quadrature (sine1) phase waveforms. The DFS is used to convert the phase value to the sine wave. The DFS Module mainly contains phase accumulator which generates the slope value, followed by complementor which generates the triangle wave values, The Multiplexor-tree generates the off-wave (one-side) values, and finally, format converter is used to generate the sine waveform. The DFS outputs (cos2 and sine2) are input to the control unit. Based on the integrator output values, the IQ output values are generated. The adder unit adds both the IQ values to generate the final GMSK modulated output.

The Channel is used to generate the noise using random sequences. The random sequences are generated using the LFSR module. For rapid implementation in Hardware, Galois field LFSR is used. The 5th order generator polynomial is used to design the Galois LFSR. The polynomial is $G(x) = x^5 + x^2 +1$. The Galois LFSR hardware design uses the 5 adders, 6 multipliers, and 6-data flip-flops. Overall two LFSR modules of the same polynomials are considered and add the two LFSR outputs to frame the random sequence. These random sequences are the same as standard AWGN generation in real time consideration. The GMSK Modulated output is XOR with channel output to produce the corrupted output values which are input to GMSK Demodulator.

The GMSK Demodulator is used to recover the similar GMSK inputs, and it mainly consists of FM demodulator using pipelined or optimized CORDIC models for IQ generation and DFS for IQ modulation, along with Differentiator and NRZ decoder. The hardware architecture of the GMSK Demodulator is represented in figure 6.

The FM demodulator receives the corrupted data (demod_in) as an input to the control unit. The pipelined or optimized CORDIC modules generate the cosine (cos1) waveform based on the phase input values. The CORDIC output is input to the DFS process, which generates the arbitrary waveform as cos2. The Delay unit is used to synchronize the DFS Modulated output with demodulated input (demod_in). The control unit is used compare the delay DFS output and demodulated input which generates the 1-bit FM demodulated output. The differentiator process the FM demodulated output using four registers subtract or. The fourth register output is subtracted from the first register output to generate the differentiator output, and it is input to the NRZ decoder. The NRZ decoder decodes the differentiator output using the delay unit and the XOR Module. The differentiator output and delayed differentiator output are inputs to the XOR module. The XOR output is inverted and generates the 1-bit GMSK demodulated output which is almost similar to the original GMSK input data sequence.

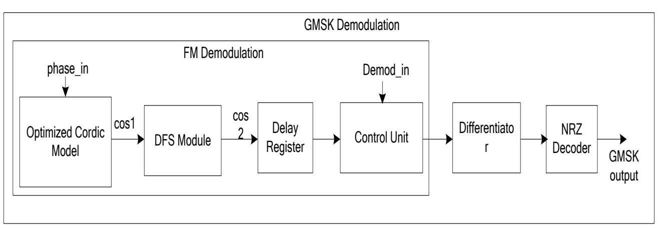


Fig. 6. Hardware Architecture of GMSK Demodulator.

V. Results and Analysis

The pipelined CORDIC and Optimized CORDIC based GMSK system results are analyzed in the below section. The two different designs are modeled over Xilinx 14.7 platform using Verilog-HDL and simulated on Modelsim6.5f and finally prototyped on Artix-7 FPGA.

The resource utilization comparison of both the pipelined CORDIC and optimized CORDIC based GMSK system are tabulated in table 2, and the graphical representation is in figure 7. The comparative results are analyzed in terms of area, time and power. The optimized CORDIC based GMSK system utilizes only 145 slice registers, working at 259.477 MHz, and utilizes 0.09W total power consumption on Artix-7 FPGA.

The optimized CORDIC based GMSK system improves the area overhead around 41.53% in slice Registers, 44.75% in slice LUTs, and 38.13% in LUT-FF pairs than pipelined CORDIC based GMSK system. The optimized CORDIC based GMSK system improves timing overhead around 38.8% in operating frequency and reduction of 2.17% in total power utilization than pipelined CORDIC based GMSK system.

The summary of the hardware synthesis results shows that the optimized CORDIC based GMSK system is better than the pipelined CORDIC based GMSK systems. The optimized CORDIC model uses Quadrature mapping along with pipeline structure, and constant tangent values are assigned directly while processing in the iteration stages. But in the Pipelined CORDIC model, the constant tangent values are assigned based on the counter method.

The hardware resource comparison of both the pipelined CORDIC and optimized CORDIC models is tabulated in table 3, and the graphical representation is in figure 8. The pipelined CORDIC Model utilizes 146 slice registers, 331 Slice LUTs and 124 LUT-FF pairs on FPGA which is quite higher than the optimized CORDIC Model. The Pipelined CORDIC operated at 157.588 MHz maximum frequency with the minimum period of 6.346ns on Artix-7 FPGA.

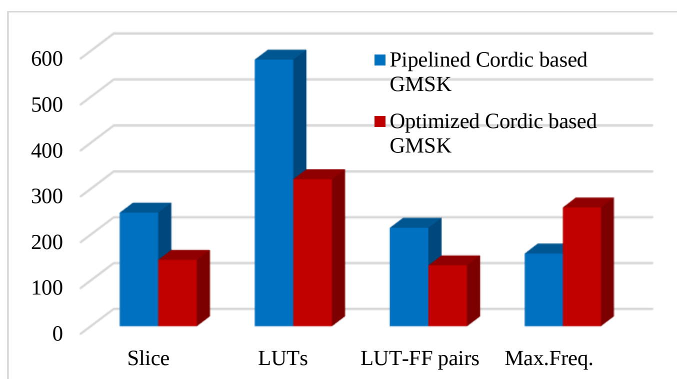


Fig. 7. Resource utilization Comparison.

TABLE. II. RESOURCE UTILIZATION OF BOTH PIPELINED AND OPTIMIZED CORDIC BASED GMSK SYSTEM

Resource Utilization	**Pipelined CORDIC based GMSK System**	**Optimized CORDIC based GMSK system**
Area		
Slice Registers	248	145
Slice LUTs	581	321
LUT-FF pairs	215	133
Time		
Minimum Period (ns)	6.294	3.854
Max.Frequency (MHZ)	158.887	259.477
Power		
Dynamic power (W)	0.01	0.008
Total power (W)	0.092	0.09

TABLE. III. RESOURCE UTILIZATION OF BOTH PIPELINED AND OPTIMIZED CORDIC MODELS

Resource Utilization	**Pipelined CORDIC Model**	**Optimized CORDIC Model**
Area		
Slice Registers	146	98
Slice LUTs	331	114
LUT-FF pairs	124	81
Time		
Minimum Period (ns)	6.346	1.962
Max.Frequency (MHZ)	157.588	509.71
Speed		
Latency(Clock cycles)	8	7
Throughput(Mbps)	157.588	582.52

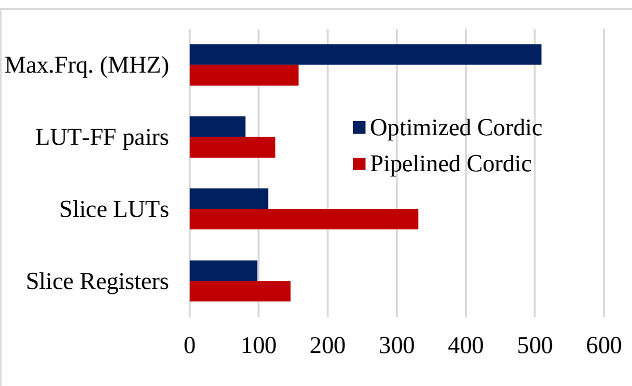


Fig. 8. Resource Comparison of Both Pipelined and Optimized CORDIC Models.

The optimized CORDIC model improves the area overhead around 32.87% in slice Registers, 65.55% in slice LUTs, and 34.67% in LUT-FF pairs than pipelined CORDIC based GMSK system. The optimized CORDIC based GMSK system improves timing overhead around 69 % in operating frequency than the pipelined CORDIC based GMSK system.

The speed comparison optimized CORDIC based GMSK system in terms of latency and throughput is improved around 12.5% and 72.89% respectively than pipelined CORDIC based GMSK system.

VI. CONCLUSION AND FUTURE WORK

The proposed GMSK system is designed using pipelined CORDIC and optimized CORDIC individually. The CORDIC Model is used to generate the IQ waveforms with low latency, which improves the speed of the GMSK system in real time scenarios. The pipelined CORDIC models is designed using shift and add method with pipeline registers along with tangent values used by the counter method. The optimized CORDIC model is designed using quadrant mapping and pipeline structure. Both the CORDIC models are designed for six stages. The hardware architecture of the GMSK system is implemented on Artix-7 FPGA with prototyping. The performance analysis of two CORDIC models based GMSK system is synthesized and summarize the hardware constraints like area, time and power. The optimized CORDIC based GMSK system improves the area (slices) around 41.53%, 38.8% in operating frequency and 2.17% in power utilization than pipelined CORDIC based GMSK system. Similarly, the optimized CORDIC model utilizes fewer resources in terms of the area (slices) around 32.87%, operating frequency around 69%, 12.5% in latency and 72.89% in Throughput than pipelined CORDIC model. In the future, adopt the proposed GMSK system in GSM standard for real-time usage and also for further spectral efficiency enhancements.

References

[1] Bax, Walt T., and Miles A. Copeland. "A GMSK modulator using a/spl Delta//spl Sigma/frequency discriminator-based synthesizer." IEEE Journal of Solid-State Circuits 36.8 (2001): 1218-1227.

[2] Gudovskiy, Denis A., Lichung Chu, and Shinhaeng Lee. "A novel nondata-aided synchronization algorithm for MSK-type-modulated signals." IEEE Communications Letters 19, no. 9 (2015): 1552-1555.

[3] Shah, Avni, and Sunil B. Somani. "An FPGA based generalized parametrizable modulator." In 2012 12th International Conference on Hybrid Intelligent Systems (HIS), pp. 402-407. IEEE, 2012.

[4] Sen, Du, Yang Yuhong, Cui Penghui, and Guo Bin. "Research on the separability of GMSK mixed signals based on modulation parameters." In 2015 IEEE 5th International Conference on Electronics Information and Emergency Communication, pp. 175-177. IEEE, 2015.

[5] Xiaopeng, Gao, and Dong Deping. "Accelerate GMSK demodulation using GPU." System Science, Engineering Design and Manufacturing Informatization (ICSEM), 2011 International Conference on. Vol. 2. IEEE, 2011.

[6] Lee, Juan Felipe Medina, Juan Felipe Patarroyo Montenegro, Catalina Muñoz Morales, Alexander López Parrado, and Juan José Giraldo Gutiérrez. "Implementation of a gmsk communication system on fpga." In 2011 IEEE Second Latin American Symposium on Circuits and Systems (LASCAS), pp. 1-4. IEEE, 2011.

[7] Liu, Shugang, and Hongli Liu. "A linear approximation of GMSK modulation for GSM/EDGE mobile communications." In 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1-4. IEEE, 2008.

[8] M. A. Adarsh, K. Adithya and A. Nimbal, "FPGA Implementation of Space-Based AIS," 2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, 2018, pp. 1-5.

[9] S. P. Majumder and M. A. Jumana, "Performance analysis of a space-frequency block coded OFDM wireless communication system with MSK and GMSK modulation," 2014 International Conference on Electrical Engineering and Information & Communication Technology, Dhaka, 2014, pp. 1-5.

[10] S. R. Udawant and S. S. Magar, "Digital image processing by using GMSK," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, 2016, pp. 695-698.

[11] Munir, Achmad, Nazmi Febrian, and Antrisha Daneraici Setiawan. "Signal analysis of GMSK modulation-based CubeSat automatic identification system receiver." In 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), pp. 439-442. IEEE, 2016.

[12] Poletacv, Alexander S., Alexander G. Chensky, and Dmitrii A. Tokmachev. "A Method for Central Frequency Phase Variations Measurements of GMSK Modulated Signals." In 2018 XIV International Scientific-Technical Conference on Actual Problems of Electronic Instrument Engineering (APEIE), pp. 101-104. IEEE, 2018.

[13] Rakesh, N. "Performance Analysis of BER for GMSK Signal for GSM Frequency Transmitted Through AWGN Channel." In International Conference on Intelligent Data Communication Technologies and Internet of Things, pp. 619-626. Springer, Cham, 2018.

[14] George, Gejo, and Srija Unnikrishnan. "Performance Evaluation of GMSK for Digital Mobile Radio Telephony." International Conference on Advances in Computing, Communication, and Control. Springer, Berlin, Heidelberg, 2013.

[15] Ghnimi, Said, AdnenRajhi, and Ali Gharsallah. "BER performance of GMSK modulation under radio mobile propagation environments." Mediterranean Microwave Symposium (MMS), 2011 11th. IEEE, 2011.

[16] Babu, KM Nitin, and K. K. Vinaymurthi. "GMSK modulator for GSM system, an economical implementation on FPGA." Communications and Signal Processing (ICCSP), 2011 International Conference on. IEEE, 2011.

[17] Gupta, Subhashini, Vikas Bhatia, and L. C. Mangal. "An efficient FPGA implementation of GMSK (BT= 0.3) transceiver with non-coherent sequence detection for tactical V/UHF waveforms." In 2012 National Conference on Communications (NCC), pp. 1-5. IEEE, 2012.

[18] Jhaidri, Mohamed Amine, Christophe Laot, and Alain Thomas. "Nonlinear analysis of GMSK carrier phase recovery loop." In 2016 International Symposium on Signal, Image, Video, and Communications (ISIVC), pp. 230-235. IEEE, 2016.

[19] L. Kumar, D. K. Mittal, and R. Shrestha, "VLSI-design and FPGA-implementation of GMSK-demodulator architecture using CORDIC engine for low-power application," 2016 IEEE Annual India Conference (INDICON), Bangalore, 2016, pp. 1-6.

[20] Li, Bingyi, Linlin Fang, Yizhuang Xie, He Chen, and Liang Chen. "A unified reconfigurable floating-point arithmetic architecture based on CORDIC algorithm." In 2017 International Conference on Field Programmable Technology (ICFPT), pp. 301-302. IEEE, 2017.

[21] Hoang, Trong-Thuc, Duc-Hung Le, and Cong-Kha Pham. "VLSI Design of Floating-Point Twiddle Factor Using Adaptive CORDIC on Various Iteration Limitations." In 2018 IEEE 12th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), pp. 225-232. IEEE, 2018.

[22] Madi, A. Ait, and A. Addaim. "Optimized Method for Sine and Cosine Hardware Implementation Generator, using CORDIC Algorithm." International Journal of Applied Engineering Research 13, no. 1 (2018): 21-29.

[23] Changela, Ankur, Mazad Zaveri, and Anurag Lakhlani. "FPGA implementation of asynchronous mousetrap pipelined radix-2 cordic algorithm." 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). IEEE, 2018.

HybridFatigue: A Real-time Driver Drowsiness Detection using Hybrid Features and Transfer Learning

HybridFatigue: Driver Fatigue detection by Abbas Q

Qaisar Abbas
College of Computer and Information Sciences
Al Imam Muhammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

Abstract*—Road accidents mainly caused by the state of driver drowsiness. Detection of driver drowsiness (DDD) or fatigue is an important and challenging task to save road-side accidents. To help reduce the mortality rate, the "HybridFatigue" DDD system was proposed. This HybridFatigue system is based on integrating visual features through PERCLOS measure and non-visual features by heart-beat (ECG) sensors. A hybrid system was implemented to combine both visual and non-visual features. Those hybrid features have been extracted and classified as driver fatigue by advanced deep-learning-based architectures in real-time. A multi-layer based transfer learning approach by using a convolutional neural network (CNN) and deep-belief network (DBN) was used to detect driver fatigue from hybrid features. To solve night-time driving and to get accurate results, the ECG sensors were utilized on steering by analyzing heartbeat signals in case if the camera is not enough to get facial features. Also to solve the accurate detection of center head-position of drivers, two-cameras were mounted instead of a single camera. As a result, a new HybridFatigue system was proposed to get high accuracy of driver's fatigue. To train and test this HybridFatigue system, three online datasets were used. Compare to state-of-the-art DDD system, the HybridFatigue system is outperformed. On average, the HybridFatigue system achieved 94.5% detection accuracy on 4250 images when tested on different subjects in the variable environment. The experimental results indicate that the HybridFatigue system can be utilized to decrease accidents.

Keywords—Driver fatigue; image processing; deep learning; transfer learning; convolutional neural network; deep belief network

I. Introduction

Driver fatigue detection (DDD) is one of the main challenging tasks to save road accidents especially in the case of Saudi Arabia where highway driving is a very common way of traveling. It was noticed that the 30% majority of road accidents caused by driver fatigue [1]. According to WHO in the year 2015 [34], road accidents are rapidly increasing. Since the drivers are often tired due to environmental conditions or mental stress and if preventive actions did not perform before an accident occur then it would cause serious deaths. A driver drowsiness detection (DDD) [5] systems are always played a vital role to prevent road accidents. In advance, the development of DDD systems is always in practice to warn drivers. This state is also known as drowsiness that causes insufficient attention to the road. At present, researchers are focusing more on developing an efficient and cost-effective solution to predict the driver's drowsiness.

According to literature, artificial intelligence and computer vision algorithms are mostly utilized in the past systems to detect driver fatigue. In practice, there are several vision-based techniques proposed in the past to detect fatigue. Due to the change of weather conditions, environmental parameters and long-drive are main caused observe in the past. Although, state-of-the-art DDD systems are considered effective and reliable during day-time but provided subtle results during nighttime driving. In previous works, we observed that there are several problems presented. Those problems are natural light, head center-off position and the detection of eyes due to large sunglasses were not perfect. Moreover, most modern methods tested on video data where people simulate fatigue behavior but less focused on real-time detection of driver fatigue. They did not experiment with real-time data, where people are naturally stressed. Since the last years, many researchers were working on developing DDD systems using many diverse [3] techniques. The driver's attentiveness can also be driver-driven with a focus on vehicle manner. Other technical methods are based on the driver's state. The best ways for correct and relatively accurate detection depend based on human drivers such as brain waves and heartbeats also rate breathing. A visual example of these DDD systems is displayed in Fig.1.

Among all those state-of-the-art DDD methods, the authors developed based on physiological phenomena that are well-thought-out as the most correct process to predict DDD driver's state. In fact, these methods are considered as accurate but all inputs should be physically connected to the driver's body. As a result, this process caused the driver to be disturbed and distracted. Besides, long driving may cause sweating on sensors (especially in KSA where the temperature is relatively high), which reduces their accuracy to monitor accurately. The second method is well suited for driving conditions in the real world because it may be non-intrusive to use camcorders to detect changes. For example, there was a smart car project, where many ECG sensors, PV, breathing, and skin behavior [2], [26] were deployed in the car. Moreover, the visual information was merged to confirm the output of the sensor. In

Toyota (ASV), the driver must wear a wrist strap to measure the heart rate. Other techniques control eyes and eye movements using a helmet or special contact lenses [4]. Those DDD methods, although less invasive, remain unacceptable in practice. Driver caution can also be exercised through indirect vehicle behaviors such as side position, wheel movements and time required for the vehicle line.

Various methodological techniques for the detection and prediction of driver fatigue have been developed through advanced image processing, machine learning, and computational intelligence algorithms. The purpose of this study is to first investigate the underlying intelligent algorithms that are applied to detect driver fatigue and then used advanced techniques to accurately detect driver fatigue suitable in the day and night-time driving. Although there are several DDD systems to address this issue according to our limited knowledge, none of them did use the multi-cam approach with hybrid features using two-cameras and heart-beat sensors. There is a recent survey article for real-time detection of driver drowsiness but that study suggested that there are still many improvements are needed in this domain. The recent trends show that the researcher is developing a hybrid system to combine visual and non-visual based features to get higher accuracy of predicting. Accordingly, in this paper, a novel HybridFatigue system was developed to address all state-of-the-art issues.

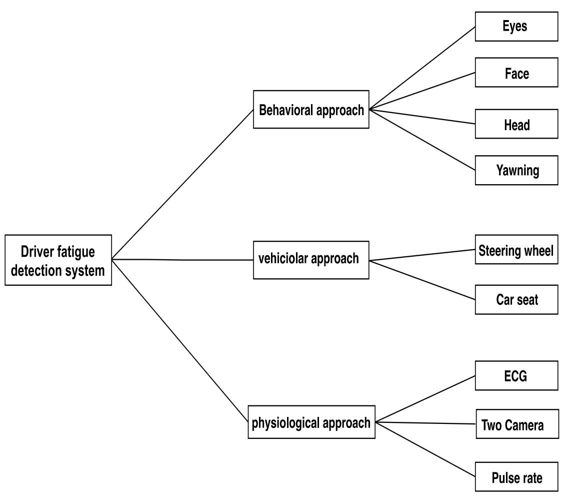


Fig. 1. A Visual Example of Driver Fatigue Detection (DDD) Systems based on different Apporaches.

II. Background

Several intelligent DDD-systems are examined in the past studies to enable the identification of driver fatigue, which enables to build the future systems. Many conventional algorithms such as Fuzzy logic, statistical models and decision trees as well as modern deep learning algorithms are described and compared. Those intelligent algorithms for DDD systems are widely utilized [6] in the past automatic systems to predict driver fatigue. Therefore, these algorithms are mainly selected for introducing and overcome the application of these methodologies. In this paper, visual features are extracted using multi-cams, heart-beats are measured through ECG sensors and those hybrid features are classified using transfer deep-learning approach. Those algorithms are usually implemented and tested through a set of rules or models to recognized drivers state during driving in various conditions. The algorithms are mainly mentioned to recognize visual and non-visual features. In addition to this, the authors have also combined one and another intelligent algorithm to get sufficient classification accuracy. In some hybrid systems, the authors have combined visual and non-visual based approached together for getting higher accuracy compared to non-hybrid methods as described in the subsequent subsection.

All the state-of-the-art classification techniques for DDD systems are error-prone as described below. Hence, they have some limitations and problems regarding their use in real-life scenarios. In practice, the authors focused more on developing DDD assistance systems and they aimed at helping drivers for detecting their moments of distraction and generate warning alarm. In the past automatic fatigue detection systems, the authors are developing hybrid systems through the outputs of diverse sensors, such as camera, vehicle sensors, and body sensors along with facial features. These hybrid systems would definitely provide a more robust and reliable decision for the prediction of driver drowsiness. To increase the robustness of hybrid systems [7], it might be possible to assigned weight to each sensor after training the machine classifier. Moreover, the redundant hardware sensors or camera should be used to avoid hardware-sensing failure. However, this process will increase the cost of hardware. Despite these facts, the selection of machine learning algorithms is also an important task. The time and space complexity always plays a tradeoff for the deployment of these driver inattention monitoring systems. Those issues are addressed in this HybridFatigue system.

III. Related Work

The driver fatigue detection (DDD) system can be divided into two classes. First, the driver's fatigue is detected and distracted only by treating the eyes areas. There are several types of research papers, which are based on this technique. The driver's eyes are the main components of the human face that can be used to track the driver's fatigue. In addition to this, treating the eye regions rather than treating the facial-area has less arithmetic complexity. In the second category, fatigue and distraction are not only detected from the eyes but also from other areas of the face and head. Moreover, many authors are also detecting other regions including yawning [19] and head nodding. In the past, many of the research presented by the authors focused on this problem. The system of detection fatigue during driving under the influence of drowsiness, facial and eye analysis system used to detect the stress of the driver using the analysis of the mouth and yawning [9]. A critical issue to be discovered and addressed is the question of how to detect accurately and efficiently in the initial stages in case of driver's head out of focus, eyes glasses and face occlusion problems.

Several methods have been proposed to detect the iris of the driver. Some of them utilized Infrared cameras [8] to simplify the iris detection problem. When infrared light is existing on the camera's center, the iris appears as bright spots due to the reflection of the blood-rich retina. Thus, the IR approach has acquired a very common list of eye detectors and functions to monitor the driver's attention [10]. Although infrared-based methods work well at night, these methods are often dysfunctional during the day because of the appearance of sunlight. Moreover, when the driver does not look at the camera center the pupils' it becomes hard to detect. Another drawback in infrared-based methods is the need to install an infrared LED's setting. Compared with infrared cameras, the CMOS and CCD cameras are passive, which means there is no infrared. The long-term effect of infrared radiation should be studied to ensure that there is no risk to eye health [11].

In [9,10], the authors developed a system to monitor the fatigue by alerting. The main idea is to put two cameras from different angles. The center of the camera corresponds to two overlapping rings, where several LEDs are equally distributed along with their surroundings. One camera has a wide range of vision to track the head, and the other has a narrow vision that focuses on eye detection. In [12], the small-square integration model was described to predict the trend towards drowsiness. They proposed a new method for modeling driver sleepiness with many of the features based on PLSR-based eyelid movement. To address the problem of strong linear relationships between the blinking of the eyes, they used partial least squares (PLSR) to predict drowsiness. The accuracy and predictive power of the generated sample are validated, indicating that it provides a new way to combine multiple features together to enhance the ability to detect and predict drowsiness.

The method of the Viola-Jones method [13] for detecting objects capable of processing images so quickly. Many researchers are utilized in this method that has a high rate of detection achieved. Moreover, this method is useful for making quick assessments of features and reducing the complexity of feature discovery for each frame. Afterward, the authors applied a machine-learning algorithm to choose a small number of features by using the AdaBoost algorithm. Also, they combined classifiers in a cascade structure. The algorithm should avoid false positive and negative values to achieve performance through which it can work. To process it, we have to require a precise numerical description of what distinguishes human faces from other objects. The recent literature review suggested [14] that there were many authors who adopted different procedures and techniques to solve the problem of driver drowsiness. However still, fatigue detection is a difficult and challenging task in the domain of computer vision. In the view of the computer vision domain, visual attention and drowsiness were detected from the live video sequence by Smith et al., in Ref. [15]. In that paper, the authors determine the state of fatigue by defining visual attention. In practice, visual attention is a process of the visually looking state of fatigue contrary to sleeping or similar conditions. For the driver's fatigue detection, those parameters should be also included to monitor the driver. To detect the visual attention parameter in the past studies, the authors are discussing that it is very problematic but it can be measured through color statistics of drivers' head and facial features. The authors tested different parameters to develop this system for driver's fatigue detection such as eye month detection, head rotation, detection of eye-blinking and eye closure in different viewing directions. Despite these visual-features, the authors also utilized sensors-based methods to detect driver drowsiness through real-time video sequences. However, to limit the scope and complexity of this review article, the driver's fatigue detection through visual-features based detection methods is discussed in detail. The subsequent paragraphs are detailed to describe those developed systems.

On 7100 frames, they achieved significant results with different illumination levels in the daytime along with 8 different drivers. A fuzzy classifier was developed in [16] by Bergasa et al., to identify the driver's vigilance in the real-time video sequence. In that research study, the authors calculated six parameters such as Percent Eye closure (PERCLOS), eye closure duration, blinks frequency, nodding frequency, face position, and fixed gaze. These six parameters are then used to recognize the driver's fatigue. To recognize the driver's fatigue, these six parameters are submitted to a fuzzy classifier for automatically determining the level of passiveness in the course of driving time. The authors tested and evaluated this system on day-night driving conditions on the motorway and achieved a near 100% accuracy. It was also noticed that this system did not perform well during the driver's wearing glasses. Furthermore, there is another system for detecting driver's vigilance was industrialized in [17, 18]. Instead of just utilizing the software-based programming techniques, the authors also implemented those techniques in hardware for real-time image acquisition, controlling the illuminator and the alarm system. They mounted the two CCD cameras on the car's dashboard. The first camera is focused on the driver's eyes, while another one is on the head movements of the drivers during driving time. This presented system was tested on different ethnic groups with different age groups with wearing glasses under different lighting conditions.

Similarly in [25], the authors utilized non-visual features such as physiological EEG and EOG signals. They used EOG signals from the forehead region of the face. The authors did experiments on 21 drivers and the system was developed through stack-based autoencoders (SAEs). On average, the authors achieved 0.85 correlation and root mean square error of 0.09. Completely in [27], the authors used visual features to develop the DDD system. A single-camera was mounted on a car dashboard to extract facial landmarks and then two SAEs were utilized to train the network for classification tasks. On average, they achieved 96.2% detection accuracy on the privately collected dataset from the driving environment. Whereas in [28], the authors used non-visuals EEG signals to develop the DDD system. The EEG signals were also used in paper [20] to detect fatigue. In that study, the authors used a combined approach of SVM and deep-learning algorithms. The deep belief network (DBN) based DDD system obtained 91.10 % of sensitivity and 55.48 % of specificity. The same approach was also utilized in [29] by using the CNN model on EEG signals. The authors reported that deep-learning methods achieved better accuracy compare to traditional machine-learning algorithms [21-24].

The convolutional neural network (CNN) [30] model was also used to develop the DDD system. In this study, a CNN modal was proposed to extract visual features from the face using a single camera. They tested this CNN model on 4,846 real eye images in the Closed Eyes in the Wild (CEW) database. Commonly used CNN models are used on the same data to compare performances of the prepared model. They obtained 96.5% detection accuracy. Whereas in [31], the authors used non-visual features such as ECG with SAEs deep-learning-based architecture. By using SAEs and ECG heart rate variability (HRV), they obtained a 90% detection rate. A different approach was adopted in [32] to detect driver distraction using four deep-learning-based architectures such as VGG-16, AlexNet, GoogleNet and residual neural network (RNN). Among these four architectures, the authors noticed that GoogleNet is the best candidate to detect driver distraction. In addition to this, a pre-trained CNN model on visual features with RNN-LSTM was also used in [33] to develop the DDD system.

Compare to the above deep-learning models, the authors in [35] used a CNN model with a transfer learning technique to develop the DDD system. On 5.5% misclassification accuracy was reported on a pre-trained model of AlexNet through visual features. Also in [36], the authors deployed hybrid features combined visual and non-visual physiological singles with the transfer-learning approach. Similarly in [37], the authors used pre-trained Alex-Net and VGGNet on hybrid features. The authors obtained an average accuracy score was 94.31%.

IV. Methodology

The systematic flow diagram of all steps of the proposed HybridFatigue system is displayed in Fig. 2. It observed from past studies that the hybrid features [32] along with transfer learning outperforms [36, 37] compare to other approaches for the development of real-time DDD systems. Therefore in this paper, hybrid features were used but extracted through a multi-cams approach that differs this HybridFatigue system to others. Overall the proposed HybridFatigue system is developed based on four main stages that are mentioned below. To develop this system, the first step is to face and then use 68-different landmarks point to track driver fatigue from the background video frames. After detecting and segmenting the landmarks points from two-mounted cameras on vehicle, the next step is to extract different visual features along with ECG non-visual features. After collecting those hybrid parameters thorough the CNN model at the collection site, this intelligent data is then passed on to automated and operator-assisted applications, which is developed through the DBN model for classification. To train the CNN model for transfer learning, the three different datasets were utilized as below.

A. Acquisition of Datasets

Three different online sources are used to test and develop the HybridFatigue DDD system. On total, there are 4250 images gathered from these three sources. These three public datasets are described in the subsequent paragraphs.

Columbia gaze dataset (CAVE-DB) [38] was collected to detect various gaze directions and head positions in a different environment. In this dataset, the authors have performed experiments on 56 people and gathered 5880 images over varying head poses and gaze directions. It contains a total of 5,880 high-resolution images of size (5,184 x 3,456) in pixels. In the case of driver's fatigue, the authors utilized head and gaze direction tracking to train their automatic systems. However, this dataset cannot be used to detect driver fatigue in a real-time environment. Also, it can be used as a best candidate for training of classifier. There are 2850 images utilized from CAVE-DB dataset. A visual example of this dataset was displayed in Fig. 3.

Multimodality Drowsiness Database is also called DROZY [39] was collected from laboratory signal and Image Exploitation (INTELSIG). Each image in this DROZY dataset was represented 14 young, healthy subjects (3 males, 11 females). These subjects in the DROZY dataset recorded 10-min hypovigilance tests that can be used to test and train the network. A visual example of the DROZY dataset is represented in Fig. 4. From DROZY dataset, the 400 images with different drossy conditions utilized.

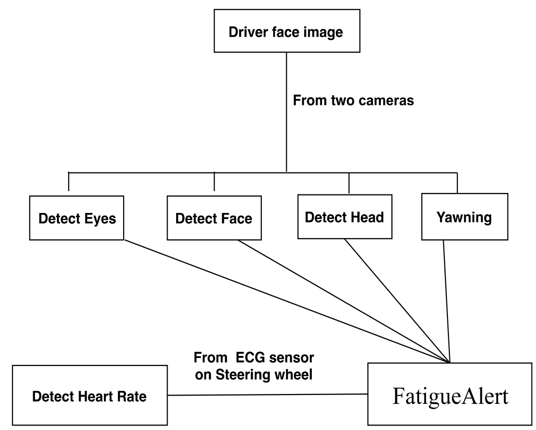


Fig. 2. A Systematic Diagram of Develop Systems of HybridFatigue Algorithm.

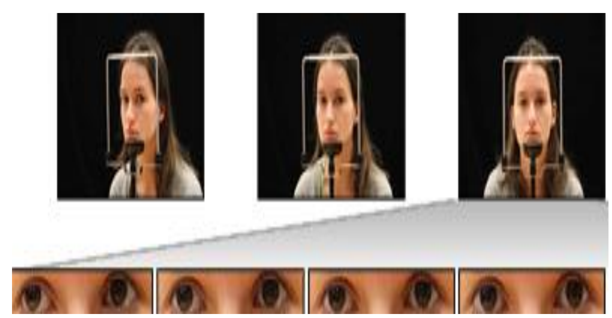


Fig. 3. An Example of CAVE-DB Dataset utilized in many State-of-the-Art Driver Fatigue Detection Systems.

Fig. 4. An Example of CEW Dataset utilized in many State-of-the-Art Driver Fatigue Detection Systems.

Fig. 5. A Visual Example of DROWSY Dataset utilized in many State-of-the-Art Driver Fatigue Detection Systems.

Many driver fatigue algorithms also utilized closed eye wild (CEW) [40] datasets to investigate the performance of eye detection algorithms. This CEW dataset contains 2423 different subjects with different eye open or closed status. Each image in the CEW dataset is of size (100×100) pixels and then extract eye patches of 24×24 centered at the localized eye position. A visual example of the CEW dataset was displayed in Fig. 5. On total, 1000 images are used as a training case.

To train the convolutional neural network (CNN) layers from these datasets, the region-of-interest (ROIs) are extracted from each central position of the driver fatigue image. These ROIs image features are used as to get transform visual features. These visual high-level features are defined from these ROI image to effectively train the multilayers of the CNN model [41].

B. Pretrain CNN Model

Convolutional neural network (CNN) model [41, 42] is widely utilized in several computer vision applications to detect features in different domains. In general, this CNN model is the best candidate of deep learning algorithms to detect the pixel-level features using convolutional filters in different layers. In the end, those features are classified through a softmax linear classifier. In the state-of-the-art DDD systems, the authors reported significantly higher performance to compare to manually tune learning algorithms. As a result, the CNN model is provided deep-invariant features than manually extracted features using digital image processing. If we need to extract and define optimize features then those deep-invariant features are the best candidates for classification tasks. In general, the convolutional neural network (CNN) model was developed through multi-layer approach. This multilayer CNN model is used to define the features-map layer. In fact, the feature-map layer was generated through the output layer. However, this features-map is not optimize as generated by convolutional filters. So, it is not applicable to use in a real-time processing due to time complexity. Accordingly to get optimize features-map, a pre-train CNN model was used that is type of transfer learning approach.

A pre-training strategy was used to guide CNN for defining the features from driver fatigue images. Three mainly groups were used to train the CNN model such as fatigue, yawn and normal including glasses. Those images are captured from three different online datasets such as CAVE-DB, CEW, and Drowsy. To develop a pre-train CNN model, top-layer of the CNN model was utilized to capture features and later on those features are transformed into weights. These weights are then utilized as to train the next layer of the CNN. To achieve invariance results, the weight-sharing and max-pooling techniques are integrated in this pre-train CNN model. Weight-sharing is also helped to reduce over-fitting results and also provided reduced train parameters.

The central-titled faces are detected by using two cameras mounted on the front dash-board and right-side of the vehicle in case a central-titled camera unable to detect the driver's face. To get accurate face features, four multi-layers of the CNN model have been trained from three datasets such as CAVE-DB, CEW and Drowsy consist of a total of 4250 images. To develop the HybridFatigue system, the high-level features from three different categories are extracted through a pre-trained CNN architecture. To select the most prominent features, the convolutional and max-pooling layers are added to this multi-layer CNN architecture. Afterward, the deep belief network (DBN) two fully connected layers are added to the architecture to get optimize features for detecting three types of faces such as drowsy, sleepy and normal conditions. In case of DBN model, unsupervised architecture using RBM multi-layer was used to fine-tune the classification model.

Visual features and non-visual features are combined into a single vector, which are extracted from each video frame using four-layers a pre-train CNN model. Those hybrid features are extracted from each training dataset of 4250 images. Moreover, a single feature map was generated from every single image, which is convolved with a Gaussian mask. The first layer of the convolution neural network (CNN) model was generated by following the DBN and max-pooling concepts. To develop this pre-train CNN model, the features are obtained at the first layer are additionally learned through the next three CNN network layers. Training phase of this CNN model with features is similar to a pre-training step. Training process is similar to the pre-training step but now the training of the proposed CNN network is performed with features from the real-time input frame. At the end, the combined features are used to train the multi-layer network architecture of combined CNN and DBN techniques. The final decision is based on DBN model based on three decision classes as show in Fig. 6.

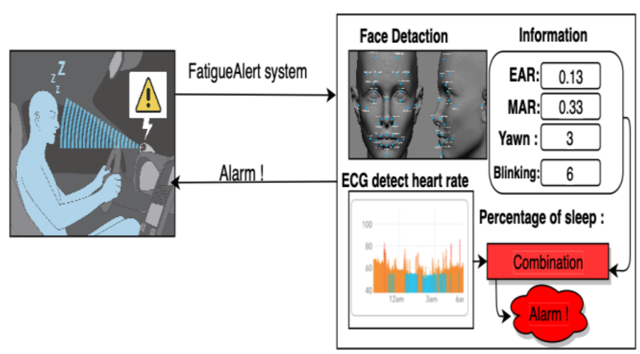


Fig. 6. A Visual Example of Built HybridFatigue Detection System using Multi-Cams Approach and Hybrid Visual and Non-Visual Features.

C. Dirver's Fatigue Prediction

Hybrid features are extracted and recognized through the pre-train CNN model and deep-belief network (DBN) multi-layer architectures. PERCLOS measure and ECG signals are both integrated into the feature vector for driver fatigue detection (DDD) tasks. A pre-train CNN model was trained on a set of 42, 50 fatigue images. A HybridFatigue detection system is detect driver fatigue into three classes through DBN classifier. This pre-train model is also capable to detect driver faces from multi-cams. Later on, a DBN model was integrated to predict driver fatigue. In this HybridFatigue system, the last two-layers are fully-connected to classify driver fatigue into three classes. The last two layers and input to this layer is a feature-map to the DBN classifier. A weighted-matrix was obtained after many times iteration using DBN model. As a result, the driver fatigue is detected through multiplying the weighted with the features and it is obtained through Eq. (1).

$$Fatigue_{i=1,2,3}^{n} = \max_{f}(W_L\, x; 0) + 1/n \sum_{i}^{n} MAR \quad (1)$$

From Eq. (1), it is cleared that the HybridFatigue detection system used w parameter to detect weights capture from DBN classifier including three fatigue classes. Also, the x parameter is used represent high-level visual features that is extracted from each frame and also optimized using a well-trained CNN model. In this equation, MAR is parameter is also added to calculate heart variable signals during sleep and finally, the decision is based on all together parameters.

V. Experimental Results

HybridFatigue detection was tested and evaluated on three different datasets such as CAVE-DB, CEW and Drowsy consist of a total of 4250 images. The HybridFatigue detection system was developed in this paper without performing any pre- or post-image processing techniques so that it can be used as real-time environment. Compare to state-of-the-art DDD system, an improved system was developed in this paper, which is based on a pre-train convolutional neural network (CNN) and deep-belief network (DBN). The combination of pre-train CNN and DBN multilayer architecture are type of transfer learning algorithms which is different than state-of-the-art methods. To develop this HybridFatigue detection system, a single RBM layer was added to extract high-level and optimize features compare to other systems. This HybridFatigue system was trained on three different datasets to improve the classification decision which is based on drowsy, yawn and normal conditions of the drivers.

The different parameters utilized to develop HybridFatigue system represented in Table I. These parameters include CPU, OS, libraries and ECG sensors with Arduino experimental board. HybridFatigue system are visually displayed in Table I. All those system components and there values are setup before to implement this HybridFatigue detection system. In this HybridFatigue detection system, CORE i5 computer with 2GB NVidia graphics GPU processing were mainly used. There are two ECG sensors are mounted on the steering wheel on both sides to get accurate heart variable readings. Those ECG sensors are connected with Arduino board and communicate through serial transmission. The complete parameters are displayed in Table I.

TABLE. I. Components and Parameters utilized for the Development of HybridFatigue System

No.	**Hardware and Computer System**	
	Components	***Parameters***
1.	CPU	Intel Core i5 3.10GHz
2.	Screen Resolution	1280 X 960
3.	Network	Ethernet Network Driver
4.	Hard Disk	250 GB
5.	Camera	720p HD video , Widescreen , Length: 4.3" / 109mm, Width: 1.75" / 44.5mm
6.	Arduino	Uno , Microcontroller: ATmega328 , Operating Voltage: 5V
7.	Sensor	ECG , The Pulse Sensor Amped is a plug-and-play heart-rate sensor for Arduino

This HybridFatigue system was developed in open source language Python with all supporting libraries related to computer vision (OpenCV), deep-learning architectures (Keras and Tensorflow) and serial communication (PySerial) tools. All those libraries are imported in Python program on 64-bit windows 10 professional. To pre-train CNN model, four multi-layer architecture was used with a batch size of 42. There were 85 epochs are setup to train this 4 layers architecture on a dataset of 4250 sample images. On approximately total, there are 120 seconds are consumed to train this architecture as a pre-train step. However on average, the HybridFatigue system took 0.345 seconds to train a single fatigue image and 0.124 seconds are taken to test an image. Table II represents the experimental results on a set of 4250 images including real-time test benchmarks. The Fatigue detection accuracy is measured to check the performance of proposed HybridFatigue system using a pre-train CNN and DBN models. The first error is measured by the detection rate while the second one is called the number of false accepts. These measures are represented in an average form in Table II.

This Table II shows the comparisons results with standard CNN model with softmax classifier, CNN and DBN compare to proposed HybridFatigue system by using a pre-train CNN and DBN models on hybrid features. The HybridFatigue detection system was evaluated on 120s video recorded during real-time driving conditions. On average, the proposed HybridFatigue detection system obtained higher 92% of detection accuracy compare to CNN of 75% and CNN&DBN of 85% value. Similarly on 135s video, the HybridFatigue is obtained higher 95% detection accuracy than CNN of 56% and CNN&DBN of 84% value. As a result, the Table II shows that the proposed HybridFatigue detection system is outperformed compare to other deep-learning methods.

Table II represents CNN and DBN deep-learning algorithms for prediction of driver fatigue based on visual and non-visual hybrid features. This table shows the performance comparison results that obtained by deep-learning methods developed in this paper. Although, there are several recent algorithms that focused only on employing the deep learning algorithms but none-of-them focused on optimize features through transfer learning. This is the main focused on this HybridFatigue detection algorithm.

TABLE I. HYBRIDFATIGUE DETECTION SYSTEM RESULTS BASED ON PRE-TRAIN CNN AND DBN MODELS.

Video ID.	HybridFatigue Performance using EAR, MAR and BPM parameters			
	Duration	*[a]CNN & [b]DBN*	*Pre-train CNN & DBN*	**Basic CNN**
1	120s	0.851	0. 920	0.745
2	135s	0.845	0.950	0.563
3	145s	0.821	0.975	0.66
Detection Accuracy		**84.50%**	**94.50%**	**65%**

[a.] CNN: Convolutional neural Network, [b.] DBN: Deep belief network

It was observed from the past DDD systems that many authors used the method of the Viola-Jones method [13] for detecting objects capable of processing images so quickly and then used classification algorithm to predict driver fatigue. The recent literature review also suggested [14] that there were many authors who adopted different procedures and techniques to solve the problem of driver drowsiness. However still, fatigue detection is a difficult and challenging task in the domain of computer vision. The reason is that the face detection is difficult where cameras are failed to detect so HybridFatigue system was developed in this paper to address those challenges.

In the view of the computer vision domain, visual attention and drowsiness were detected from the live video sequence by Smith et al., in Ref. [15]. On 7100 frames, they achieved significant results with different illumination levels in the daytime along with 8 different drivers. A fuzzy classifier was developed in [16] by Bergasa et al., to identify the driver's vigilance in the real-time video sequence. In that research study, the authors calculated six parameters such as Percent Eye closure (PERCLOS), eye closure duration, blinks frequency, nodding frequency, face position, and fixed gaze. Furthermore, there is another system for detecting driver's vigilance was industrialized in [17, 18]. They mounted the two CCD cameras on the car's dashboard. The first camera is focused on the driver's eyes, while another one is on the head movements of the drivers during driving time. This presented system was tested on different ethnic groups with different age groups with wearing glasses under different lighting conditions.

The authors utilized non-visual features such as physiological EEG and EOG signals [25]. They used EOG signals from the forehead region of the face. The authors did experiments on 21 drivers and the system was developed through stack-based autoencoders (SAEs). On average, the authors achieved 0.85 correlation and root mean square error of 0.09. Completely in [27], the authors used visual features to develop the DDD system. A single-camera was mounted on a car dashboard to extract facial landmarks and then two SAEs were utilized to train the network for classification tasks. On average, they achieved 96.2% detection accuracy on the privately collected dataset from the driving environment. The convolutional neural network (CNN) [30] model was also used to develop the DDD system. In this study, a CNN modal was proposed to extract visual features from the face using a single camera. They tested this CNN model on 4,846 real eye images in the Closed Eyes in the Wild (CEW) database. Commonly used CNN models are used on the same data to compare performances of the prepared model. They obtained a 96.5% detection accuracy. Whereas in [31], the authors used non-visual features such as ECG with SAEs deep-learning-based architecture. By using SAEs and ECG heart rate variability (HRV), they obtained a 90% detection rate. A different approach was adopted in [32] to detect driver distraction using four deep-learning-based architectures such as VGG-16, AlexNet, GoogleNet and residual neural network (RNN). To avoid these complex deep-learning architectures, the combination of CNN and DBN pre-train models are integrated together to get higher accuracy compare to the all other DDD systems.

As a result, the driver fatigue detection through deep-learning algorithms are more improved and get deeper in term of number of layers and number of processing units per layer to become in line with the latest progress in computer vision applications. So, the proposed HybridFatigue is outperformed compare to most recent developed systems for driver fatigue detection. Still, the deep-learning architectures are required to be more powerful in terms of their computational power for real-time recognition of driver fatigue. In future, there is dire need to explore some schemes to increase the multi-layer deep learning architectures in terms of computational power. All above-mention systems utilized a single camera followed by machine learning algorithms to detect driver fatigue. However, there is also a need to test driver fatigue detection system through multiple cameras. Therefor in this paper, multi-cams approach was utilized to extract visual features.

VI. CONCLUSIONS

In this paper, a novel HybridFatigue detection system was developed by integrating visual and non-visual features based on multi-cams and ECG sensors. In the first stage, the HybridFatigue system detected the driver's face which is inside the video-frame. If the face is not detected through the central camera then the second camera on the right-side pop-up to detect driver face for PERCLOS measure. At last, the face marks are calculated based on the EAR and MAR values. Those values are detected if the eyes are open or closed and the driver is yawning or not. The HybridFatigue system has been tested in a real-time environment mode during nighttime and daytime under different environmental conditions. The HybridFatigue detection system is also capable to read the heartbeat (bpm) from ECG sensor readings. At last, the HybridFatigue system is able to combine AER, MAR and BPM ratio values. Finally, the HybridFatigue system detected the drowsiness alert. The HybridFatigue system was evaluated and tested on three publically available datasets. The HybridFatigue detection system was trained on 4250 images using CNN and DBN deep-learning models. The experimental results indicate that the proposed HybridFatigue system outperformed compared to state-of-the-art DDD systems. Experimental results indicate that the HybridFatigue achieved 94.50% of accurate detection accuracy by using multi-cams approach with ECG sensors.

VII. Future Work

In fact, the HybridFatigue will be upgraded in the future to add more computational processing through cloud computing. On future, mobile application and cloud computing resources will be integrated to improve computational power and cost of HybridFatigue system. Also, three multi-cams approach will be integrated to improve HybridFatigue system in terms of detection accuracy.

Acknowledgment

This work is a part of research project funded by King Abdulaziz City for Science and Technology (KACST) under its applied research program – The National Transformation Program- grant no. (0001-008-11-17-3). The author also would like to thank my bachelor students who did partial implemented of this driver fatigue detection system.

References

[1] Connor J. The role of driver sleepiness in car crashes: a systematic review of epidemiological studies. In: Verster J., George C., editors. Sleep, sleepiness and traffic safety. Nova Science Publishers; New York: 2010.

[2] RAHIM, Herlina Abdul; DALIMI, Ahmad; JAAFAR, Haliza. Detecting drowsy driver using pulse sensor. Jurnal Teknologi, 2015.

[3] BERGASA, Luis Miguel, et al. Real-time system for monitoring driver vigilance. IEEE Transactions on Intelligent Transportation Systems, 2006.

[4] REDDY, Bhargava, et al. Real-time driver drowsiness detection for embedded system using model compression of deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017.

[5] EDWARDS, David J., et al. Evaluation of fatigue management technologies using weighted feature matrix method. 2007.

[6] SARADADEVI, Mandalapu; BAJAJ, Preeti. Driver fatigue detection using mouth and yawning analysis. International journal of Computer science and network security, 2008.

[7] ZHAO, Shuyan; GRIGAT, Rolf-Rainer. Robust eye detection under active infrared illumination. In: Proceedings of the 18th International Conference on Pattern Recognition-Volume 04. IEEE Computer Society, 2006.

[8] D. Pitts, A. Cullen, and P. Dayhew-Barker, Determination of ocular threshold levels for infrared radiation cataractogenesis: NIOSH research report, DHHS publication ; no. (NIOSH) 80-121, DHHS publication - no. (NIOSH) 80-121, 1980.

[9] Q. Ji and X. Yang, “Real-time eye, gaze, and face pose tracking for monitoring driver vigilance,” Real-Time Imaging, vol. 8, pp. 357-377, 2002.

[10] J. P. Batista, “A real-time driver visual attention monitoring system,” Proc. of Iberian Conference on Pattern Recognition and Image Analysis, vol. 3522, pp. 200-208, 2005.

[11] SU, Hong; ZHENG, Gangtie. A partial least squares regression-based fusion model for predicting the trend in drowsiness. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2008.

[12] WANG, Yi-Qing. An analysis of the Viola-Jones face detection algorithm. Image Processing On Line, 2014.

[13] ZHU, Zhiwei; FUJIMURA, Kikuo; JI, Qiang. Real-time eye detection and tracking under various light conditions. In: Proceedings of the 2002 symposium on Eye tracking research & applications. ACM, 2002.

[14] DAZA, Iván Garcia, et al. Drowsiness monitoring based on driver and driving data fusion. In: 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE, 2011.

[15] SACCO, Matthew; FARRUGIA, Reuben A. Driver fatigue monitoring system using support vector machines. In: 2012 5th International Symposium on Communications, Control and Signal Processing. IEEE, 2012.

[16] Khushaba R.N., Kodagoda S., Lal S., Dissanayake G. Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. IEEE Trans. Biomed. Eng. 2011.

[17] Liang W., Yuan J., Sun D., Lin M. Changes in physiological parameters induced by indoor simulated driving: Effect of lower body exercise at mid-term break. Sensors. 2009.

[18] SINGH, Hardeep; BHATIA, Jagjit Singh; KAUR, Jasbir. Eye tracking based driver fatigue monitoring and warning system. In: India International Conference on Power Electronics 2010 (IICPE2010). IEEE, 2011.

[19] KNAPIK, Mateusz; CYGANEK, Bogusław. Driver’s fatigue recognition based on yawn detection in thermal images. Neurocomputing, 2019.

[20] WANG, Fuwang; WANG, Hong; FU, Rongrong. Real-Time ECG-based detection of fatigue driving using sample entropy. Entropy, 2018.

[21] Facial exprssision Recognition system,

[22] https://www.exeros-technologies.com/solutions/ai-facial-recognition-system/. [Access date: 10/11/2019].

[23] Driver Fatigue monitoring System, https://www.aliexpress.com/item/stock-CE-FCC-anti-sleep-alarm-hot-Driver-Fatigue-early-MR688-Driver-Fatigue- onitor/32814066993.html. [Access Date: 12/12/2019].

[24] IBRAHIM, Masrullizam Mat, et al. Yawn analysis with mouth occlusion detection. Biomedical Signal Processing and Control, 2015.

[25] LI, Gang; CHUNG, Wan-Young. A context-aware EEG headset system for early detection of driver drowsiness. Sensors, 2015.

[26] Du, L. H., Liu, W., Zheng, W. L., & Lu, B. L. (2017, May). Detecting driving fatigue with multimodal deep learning. In 2017 8th International IEEE/EMBS Conference on Neural Engineering (NER) (pp. 74-77). IEEE.

[27] Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., & Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: A review. Computer methods and programs in biomedicine, 161, 1-13.

[28] Zhao, L., Wang, Z., Wang, X., Qi, Y., Liu, Q., & Zhang, G. (2016). Human fatigue expression recognition through image-based dynamic multi-information and bimodal deep learning. Journal of Electronic Imaging, 25(5), 053024.

[29] San, P. P., Ling, S. H., Chai, R., Tran, Y., Craig, A., & Nguyen, H. (2016, August). EEG-based driver fatigue detection using hybrid deep generic model. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 800-803). IEEE.

[30] Hajinoroozi, M., Zhang, J. M., & Huang, Y. (2017, October). Driver's fatigue prediction by deep covariance learning from EEG. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 240-245). IEEE.

[31] Tümen, V., Yıldırım, Ö., & Ergen, B. (2018, April). Detection of driver drowsiness in driving environment using deep learning methods. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-5). IEEE.

[32] Bhardwaj, R., Natrajan, P., & Balasubramanian, V. (2018, December). Study to Determine the Effectiveness of Deep Learning Classifiers for ECG Based Driver Fatigue Classification. In 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS) (pp. 98-102). IEEE.

[33] Tran, D., Do, H. M., Sheng, W., Bai, H., & Chowdhary, G. (2018). Real-time detection of distracted driving based on deep learning. IET Intelligent Transport Systems, 12(10), 1210-1219.

[34] Xiao, Z., Hu, Z., Geng, L., Zhang, F., Wu, J., & Li, Y. (2019). Fatigue driving recognition network: fatigue driving recognition via convolutional neural network and long short-term memory units. IET Intelligent Transport Systems.

[35] Jabbar, R., Al-Khalifa, K., Kharbeche, M., Alhajyaseen, W., Jafari, M., & Jiang, S. (2018). Real-time driver drowsiness detection for android application using deep neural networks techniques. Procedia computer science, 130, 400-407.

[36] Jakubowski, J., & Chmielińska, J. (2018). Detection of driver fatigue symptoms using transfer learning. Bulletin of the Polish Academy of Sciences. Technical Sciences, 66(6).

[37] Chen, L. L., Zhang, A., & Lou, X. G. (2019). Cross-subject driver status detection from physiological signals based on hybrid feature selection and transfer learning. Expert Systems with Applications.

[38] Budak, U., Bajaj, V., Akbulut, Y., Atilla, O., & Sengur, A. (2019). An Effective Hybrid Model for EEG-Based Drowsiness Detection. IEEE Sensors Journal.

[39] Smith, B.A.; Yin, Q.; Feiner, S.K.; Nayar, S.K. Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction. In Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, St. Andrews, UK, 8–11 October 2013; pp. 271–280.

[40] Massoz, Q., Langohr, T., Francois, C., Verly, J.G.: The ULG Multimodality Drowsiness Database (called DROZY) and Examples of Use, WACV (2016).

[41] F.Song, X.Tan, X.Liu and S.Chen, Eyes Closeness Detection from Still Images with Multi-scale Histograms of Principal Oriented Gradients, Pattern Recognition, 2014.

[42] Vijayan, V., & Sherly, E. (2019). Real time detection system of driver drowsiness based on representation learning using deep neural networks. Journal of Intelligent & Fuzzy Systems, (Preprint), 1-9.

[43] Abbas, Q. V-ITS: Video-based Intelligent Transportation System for Monitoring Vehicle Illegal Activities. International Journal of Advanced Computer Science and Applications(IJACSA), Volume 10 Issue 3, 2019.

An Improved Deep Learning Approach based on Variant Two-State Gated Recurrent Unit and Word Embeddings for Sentiment Classification

Muhammad Zulqarnain[1], Suhaimi Abd Ishak[2], Rozaida Ghazali[3], Nazri Mohd Nawi[4]
Muhammad Aamir[5],Yana Mazwin Mohmad Hassim[6]
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia[1, 2, 3, 5, 6]
Soft Computing and Data Mining Center (SMC), Universiti Tun Hussein Onn Malaysia[4]
86400, Parit Raja, Batu Pahat, Johor, Malaysia[1, 2, 3, 4, 5, 6]

***Abstract*—Sentiment classification is an important but challenging task in natural language processing (NLP) and has been widely used for determining the sentiment polarity from user opinions. And word embedding technique learned from a various contexts to produce same vector representations for words with same contexts and also has been extensively used for NLP tasks. Recurrent neural networks (RNNs) are common deep learning architecture that are extensively used mechanism to address the classification issue of variable-length sentences. In this paper, we analyze to investigate variant-Gated Recurrent Unit (GRU) that includes encoder method to preprocess data and improve the impact of word embedding for sentiment classification. The real contributions of this paper contain the proposal of a novel Two-State GRU, and encoder method to develop an efficient architecture namely (E-TGRU) for sentiment classification. The empirical results demonstrated that GRU model can efficiently acquire the words employment in contexts of user's opinions provided large training data. We evaluated the performance with traditional recurrent models, GRU, LSTM and Bi-LSTM two benchmark datasets, IMDB and Amazon Products Reviews respectively. Results present that: 1) proposed approach (E-TGRU) obtained higher accuracy than three state-of-the-art recurrent approaches; 2) Word2Vec is more effective in handling as word vector in sentiment classification; 3) implementing the network, an imitation strategy shows that our proposed approach is strong for text classification.**

Keywords*—*RNN; GRU; LSTM; encoder; Two-state GRU; long-term dependencies; sentence classification

I. Introduction

Automated sentiment classification is the process of extracting the opinions of various user expressed in texts, that are contains emotions or opinions behind the sentences, and identifies the positive or negative aspects of comments, which is very useful in analyzing user generated contexts. The issue of sentiment classification is a popular area in natural language processing (NLP) tasks and recently has gained a plenty of concentration. There are lots of data publically are available on various online platforms, that allow us to performs a sentiment classification for the favors of academic attention. Sentiment analysis is done at both at phrase level and document or paragraph level. Both the levels offer unique challenges and hence require different techniques to tackle them. Word embedding is a fundamental task in NLP and commonly used method to encode words into a high dimensional space. In recent years, they have taken excellent attention to the base of extract both semantic and syntactic information. The main concept of word embeddings is a long history but it has become well-known since Bingio et al. efforts [1] in which each word is shown by word vector and the concatenations of many previously word vector are used to predicts the following based on a language model. Traditionally, to represented an each word as a vector with high-dimensional sparse vector similar to the number of unique terms in the vocabulary using distributional approaches [2] such as vector context-based techniques [3] [4] and the Hyperspace Analog to Language (HAL) model [5]. Recent, a latest distributed words representations training technique, identified as word embeddings [6] [7] [8] has been established to illustrate words as low-dimensional vector for text representation of real numbers, which can effectively capturing semantic and syntactic word similarities from big datasets. These word embedding approaches has been excellently applied for several tasks included entity recognition [9], dependency parsing [10], text classification [11], and speech recognition [12].

Word embedding, the logical meaning of words that trained from specific contexts tends to produce same vector representation for word with same contexts. This technique performs better for semantic-oriented applications but its problem for sentiment classification because words with same vectors representation due to same contexts may have an contrary sentiment separation, as in the examples of happy-sad referred in [13] and positive-negative in [14]. Furthermore, most of the authors have investigated the unique features of the vector representation of words through machine learning [15]. The two well-known word embeddings representation methods of neural network language models, is called Glove word vector and TF-IDF were introduced by Pennington and Aizawa [16], [17]. However, these traditional context-based word embedding such as TF-IDF and GloVe usually unsatisfied to capturing appropriate sentiment information, which may result in words with same vectors representation having a reverse sentiment polarity. To handle this issue, in this research have suggested using Word2Vec, word embedding method to representation words in text as vectors for sentiment classification.

In recent years, deep learning architectures have gradually presented better performance in several data mining applications, such as text classification, entity recognition and sentiment analysis. These architectures effectively addressing the features representation issue because they learnt features from contextual data automatically. Between deep feed-forward neural networks, convolutional neural networks (CNNs) [18] have been presented to capture from words or phrases, and recurrent neural networks (RNNs) [19] are capable to capture temporal dependencies in sequence information and have shown strong semantic composition approaches for sentiment classification [18]. Provided large texts in social media, there is absence of features. To achieve progressively important features, we further used appropriated concept of distributed illustration of words where each input is denoted by numerous features and each feature is included in several potentially sequential inputs. In particularly, we used the pre-trained Word2Vec word embedding method [6] for distributed representation of social media.

RNNs have been extremely used in recent years for the tasks of texts classification. The key benefit of RNNs is that they can be applied to extracts temporal sequential data with variable-length, which flexibilities generates in evaluating reviews of various lengths. Recent, simplified architecture through LSTM, known as Gated Recurrent Unit (GRU), was proposed by [20]. GRUs contains fewer parameters and explained by very simple set of equations, thus need significantly less computational power. Relation among GRU and LSTM effectiveness is an initiate problem and a domain of research.

In this research, we have investigated the effectiveness of GRU for sentiment classification with distributed representation in social media. Applied variant-GRU model for the aim of preventing the issue of gradient exploding or vanishing in an existing RNNs and also overcome the deficiency of standard GRU. In this work, we conducted the experiment on two benchmark datasets, IMDB and Amazon Products Reviews. We compared the performance of proposed Encoder Two-State Gated Recurrent unit (E-TGRU) with three traditional RNNs models namely are, Gated Recurrent Unit (GRU), Long Short Term Memory (LSTM) and Bidirectional Long Short Term Memory (Bi-LSTM).

Our research work continues to further investigate standard GRU based on variant-GRU with encoder to automatic preprocessing data to provide an improved presentation of the inputs than original raw inputs. Based on previous work, our main objective is enhancing the standard GRU structure in order to increase accuracy of sentence classifications and minimize the information loss. In particularly, based on the above studied, the main contribution of this research is summarized as follows:

- Proposing a variant of GRU is included encoded gated recurrent unit (E-GRU) for sentiment analysis. This method performs automatically preprocess text data through encoder.
- Proposed a Two-State Gated Recurrent Unit (TGRU) for sentiment classification.
- To proposed Word2Vec pre-trained word embedding method is applied for sentiment classification to illustrate words as vectors in long-short term.
- Experimental results demonstrate that the (our network) namely Encoder Two-State GRU (E-TGRU) architecture performs well on both tasks to takes advantages of the encoded local features extracted and captured the long-term dependencies, and further enhance the sentiment classification performance but significantly less computational expensive than Bi-LSTM.

II. Models Descriptions

A. Recurrent Neural Network

A standard RNN is kind of artificial neural networks in which connections among the units form a bidirectional cycle, and they perform the similar task for each element in the sequence. The RNN technique is better for sequential issue especially capturing temporal information in the loop. The RNN uses recurrent hidden state whose activation that particularly dependent on the previous timestep while performing the sequential information and the current state rely on the current input. Thus, the current hidden state makes complete usage of past information. In this way, standard RNN can handle variable length and computes sequential data in dynamic processes. The architecture of an RNN is shown in Fig. 1. When provided a sequential inputs X = [x_1, x_2 ... x_t ... x_T] of length T, process output vector O_t, an RNN defines the hidden state U_o at the time t with the sequential of following equations:

$$\overrightarrow{O_t} = \varphi(\overrightarrow{W_x} x_t + \overrightarrow{U_o}\,\overrightarrow{h_{t-1}}) \tag{1}$$

$$\overleftarrow{O_t} = \varphi(\overleftarrow{W_x} x_t + \overleftarrow{U_o}\overleftarrow{h_{t+1}}) \tag{2}$$

$$O_t = [\overrightarrow{O_t} : \overleftarrow{O_t}] \tag{3}$$

where $W_x \in R^{O_h \times O_x}$ is the weights matrix connecting input layer to hidden layer, and $U_o \in R^{O_h \times O_h}$ is the hidden layers weights matrix. σ is the sigmoid activation function and W_x, U_o, are parameters of the traditional RNN.

In Fig. 1, the input unit is X_1 of time-step t, which shows the word vector of the t-*th* word in the text; h_t is the final activation state of step t; O_t shows the output at time t, the output is chosen according to the require of the network; U, W are the weights parameters of the networks are require to trained the model. However, RNNs are hard to train and suffer from vanishing and exploding gradients issue.

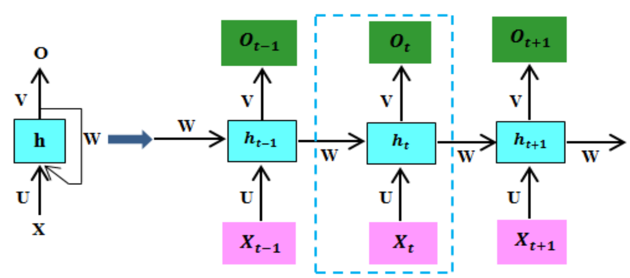


Fig. 1. Traditional Structure of RNN.

Although the RNN is very powerful when dealing with sequential problems, it is difficult to train with the gradient descent method and suffer from vanishing and exploding gradient (explosion) issue [20]. On the other hand, variants of RNN have been developed to solve the above issues, such as LSTM and GRU. Between them, GRU avoids overfitting, as well as saves training time. Therefore, GRU is adopted in our method.

B. Long Short Term Memory

Long Short-term Memory (LSTM) is one of the RNN structure that was initially proposed by [21] and has largely applied in natural language processing. LSTM consist the gated mechanism and internal cell memory that help to addresses the well-known issues relevant to vanishing gradients or exploding. The fundamental concept of “gates” used in LSTM for the aim of handling the sequence contextual data. Furthermore, a common LSTM architecture units is composed of an internal memory cell, and three gates in recurrent connection that help the model to determine how much information pass away and extracting more information with timestep. In addition, by using three gates are generally developed to handle to insert or ignore the information in the memory cell. These gates help the model to regulate how to update the current memory cell and the current hidden stateU_h.

Furthermore, this addition process is completed through three steps. First, Sigm have used as a sigmoid activation functions to process the data which require to store to the internal memory cell. Second, the tanh function is taken to achieve a vector over h_{t-1} and x_t. The final step, output gate is determined the task for choosing appropriate information from the memory cell to output.

The LSTM cells that are used in the transition functions are implemented as follows:

$$i_t = Sigm(W_{xi}x_t + U_{hi}h_{t-1} + b_i) \quad (4)$$

$$o_t = Sigm(W_{xo}x_t + U_{ho}h_{t-1} + b_o) \quad (5)$$

$$f_t = Sigm(W_{xf}x_t + U_{hf}h_{t-1} + b_f) \quad (6)$$

$$\hat{a}_t = tanh(W_{x\hat{a}}x_t + U_{h\hat{a}}h_{t-1} + b_{\hat{a}}) \quad (7)$$

$$c_t = f_t * x_{t-1} + (i_t * \hat{a}_t) \quad (8)$$

$$h_t = O_t * tanh\,(c_t) \quad (9)$$

where, Sigm is the logistic sigmoid function that produce the output between [0, 1]. The variables and bias to be calculated during the learning procedure are W_i, W_o, W_f, $W_c \in R^{m\times p}$,U_i, U_o, U_f, $U_c \in R^{m\times m}$ b_i, b_o, b_f, $b_c \in R^{m\times 1}$ and * indicates the element-wise multiplication of the two vectors. Each gate is composed out of a sigmoid layer and has the capability to remove or add information from the memory cell. At the current time t, U_h denotes hidden states, i_t, O_t, f_t indicates as a input, output and forget gate. W_i, W_o, and W_f represent the weight parameters of LSTM respectively, while b_i, b_o, b_f refers to the biases of the gates.

C. Traditional Gated Recurrent unit

Gated Recurrent Unit (GRU) is another advance kind of RNN, and simplified variation of LSTM relatively development proposed by Cho et al. in [20]. GRU is simple variation of LSTM, and that contain only two gates update gate z_t and reset gate r_t that handle the flow of information inside the unit, while without having an individual memory cells are shown in Fig. 2. GRU also illustrates the powerful capability of modeling to capturing long-term dependencies between the elements of a sequence. GRU calculates two gates, which manages the flow of information through each hidden unit. Each hidden state U_h at time t, given input x_t calculated using the following equations:

$$z_t = \sigma_g\,(W_x^z x_t) + (U_h^z h_{t-1}) \quad (10)$$

$$r_t = \sigma_g\,(W_x^r x_t) + (U_h^r h_{t-1}) \quad (11)$$

$$\tilde{h}_t = \tanh(W_x^{\tilde{h}} x_t) + U_h^{\tilde{h}}(r_t * h_{t-1}) \quad (12)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (13)$$

where z_t refers to the update gate, and r_t refers reset gate, W, U and are weights matrix and vectors. σ_g is a sigmoid activation function and tanh is the hyperbolic tangent. After developing the design of the GRU model, the learning technique for the GRU model requires to be determined. At currently, for RNNs such as GRU, become a common training method include back propagation trough time (BPTT) and real time recurrent learning (RTRL).

However, based on the recent studied there are two issues in standard GRU network. First, in data pre-processing phase much manual experience is needed to preprocess data for accurately classification and second one is high consumption of memory. Therefore, we proposed GRU variant, such as encoder E-GRU. Fig. 3 illustrates the structure of an E-GRU. The weights W, U and the activation function in the E-GRU are the similar as in the standard GRU. Furthermore, the variant E-GRU applies the encoder for automatically pre-process large amount of data. The encoder usually provides a best illustration of the input compared to original raw input, and the encoder is consistently compresses the input data in which select significant features for training.

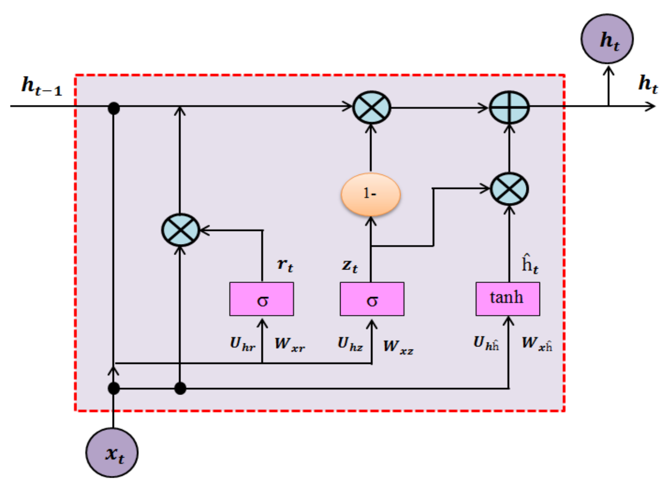


Fig. 2. GRU Architecture.

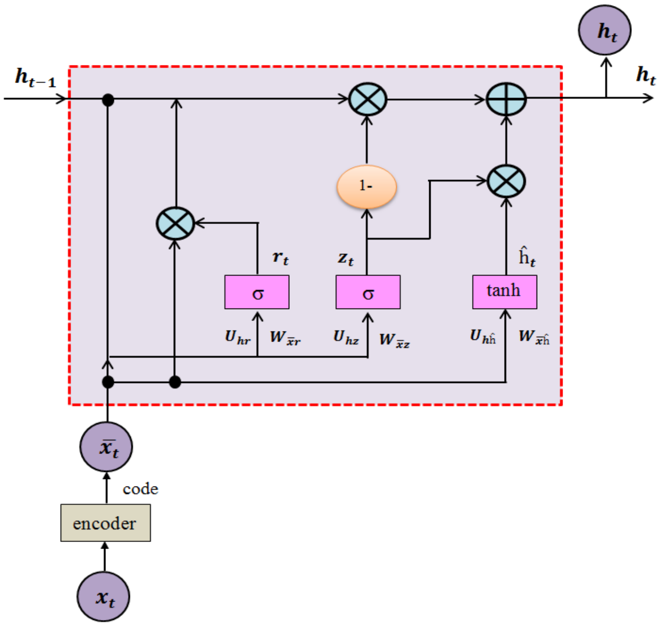


Fig. 3. Architecture of Variant E-GRU.

III. Proposed Methodology

In this section, we demonstrate the particular description of the proposed methodology architecture that contains encoder based gated recurrent unit with word embedding. The proposed methodology apply Two-State GRU which learns to extract forward and backward context features through time steps, whose outputs are then given by encoded GRU model, and finally, followed by a softmax classifier. The description of each methodology elements to solving overall sentiment classification problem are follows:

A. Gated Recurrent Unit (GRU)

It consist a gating structure and a advance type of standard RNNs. It also illustrates the powerful capability to process of sequential data and capturing long-term dependencies between the elements by preserve previous state in the internal state of model through time step t.

B. Variant Gated Recurrent Unit

The first variant consists of binary gated recurrent unit (Bin-GRU), and local feature-based GRU (LF-GRU), and so on. Zhao et al. [29] introduced LF-GRU. In LF-GRU, first it extracts local features from segment or windows of time-series data. After that, LF-GRU base model is applied to learn with average weighed features from sequential of local features. However, the above approaches are requiring more manual experience for preprocessing massive data. Therefore, we proposed encoder GRU (E-GRU). E-GRU uses the encoder for automatically preprocess data efficiently. In this way, the output of encoder (E-GRU) becomes the input of Two-State GRU appropriately. In this paper, we mostly discussed Two-State GRU and E-GRU.

A. Traditional Auto-Encoder

Auto-encoder is a kind of unsupervised artificial neural networks that learns to effectively compress and encode data. The main purpose of auto-encoder is identifying further helpful and valuable features from huge amount of dataset without application of any dimensionality reduction method. Basically, Auto encoder usually performs in two stages namely encoding and decoding. Along with encoding stage converts the input features to a new representation [22] while decoding stage tries to convert this new representation back as near as possibly to its original inputs.

B. Encoder GRU (E-GRU)

The It contain on encoder GRU that used to reduce dimensionality from input data by applying the encoder part of the auto-encoder, it gives the best representation of the inputs than original raw inputs, and then the outputs of the encoder as become the inputs of the Two-State GRU.

C. Features Extraction

In sentiment classification feature extraction technique obtain significant role for identifying relevant features from raw data. It also includes to eliminate unnecessary features [31] and maintaining important features that consider to improve the accuracy of the model.

D. Automatical Preprocessing

It presents to the automatically capturing of features from the original raw data and removes manual interference.

E. Word Embedding Layer

In sentence classification process the initial stage is pre-processing the inputs sentence and sentiment context words. To superior representation the limited contents in long and short text. In this paper, we applied the pre-trained Word2Vec [23] word embedding method in embedding layer to extract the contextual correlation between words in training data. Word2vec performs as a predictive model to train their co-occurrence vectors to extracts the correlation among the target word and the context words in a simple way: it tries to extract the relevant semantic regularities by learning with the activation of the target word and its context words. The word embedding layer of the model changes words context into real-valued features vectors that captured semantic and syntactic data. Let L ∈ R^{V*d} be the embedding query table produced by Word2vec, where *d* is the dimension of words and V is the vocabulary size. Assume that the input sentence contains of *n* words and the sentiment resource contains of m words.

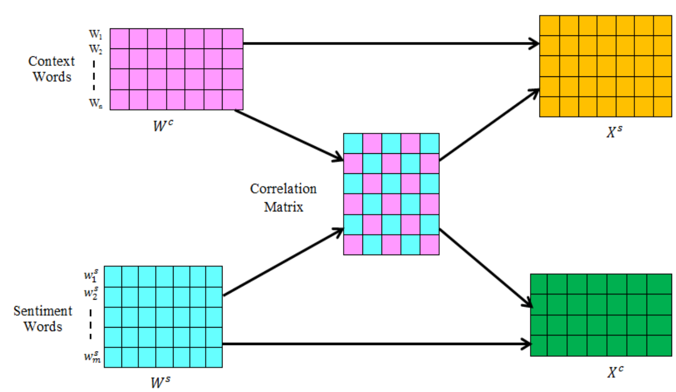


Fig. 4. Sentiment-Context Word Correlation.

The input sentence retrieves the word vectors from L and obtain a list of vectors [W_1, W_2 …, W_n] where $W_i \in R^d$ is the word vector of the i^{th} word. Similarly, the sentiment resource sequence can retrieve the word vectors from and form a list of vectors $[W_1^s, W_2^s, \ldots, W_m^s]$. In this way, we can get the matrix $W^c = [W_1, W_2, \ldots., W_n] \in R^{n*d}$ for context words and the matrix $W^s = [W_1^s, W_2^s, \ldots., W_m^s] \in R^{m*d}$ for sentiment resource words. Fig. 4 show the process is simply concatenation of all words embedding in V.

F. The TGRU Network Architecture

The GRU recurrent layer has the ability to represent sequences e.g. sentences and very useful to capturing long-term dependencies between elements of a sequence. GRU can be applied for sentiment classification in the same manner as it has been used in Cho K [24]. GRU is a recent simplified variant of LSTM, that combine the forget gate and input gate into a single new update gate, which enhance the convergence time and iteration times of model training. First, an embedding layer generated a suitable size. The embedding layer will perform to show each word by a real valued vector of similar range to the fixed dimension. These values are the weights among the embedding layer and the hidden layer on top of it. These units are not only connected to the layer below and above them but also connected to units within their own layer. At the end of the hidden layer we attain the representation of the entire sequence which can be used as input to linear model or classifier.

GRU structure are basically consists an update gate and reset gate. Reset gate (r_t) determines that how much previous memory can be ignore from previous hidden state h_{t-1}.The update gate (z_t) is determines how much previous state keep around and send among the existing state and new calculated state with parameter bias b_z. x_t as a p input vector dimension at time t , σ is the logistic sigmoid activation, and W_{zx} (q × p matrix), U_{zh} (q × q matrix), b_z (q×1 vector) are determined size parameters which are common through an whole model.

$$z_t = \sigma(W_{zx}x_t) + U_{zh}h_{t-1} + b_z) \quad (14)$$

where σ is the sigmoid activation applied for binary classification in the dense layer (output layer), and the value range of each element in the update gate z_t are [0, 1].

The reset gate r_t is calculated similarly to the update gate but with changed weights value: W_{rx}(q × p matrix), U_{rh} (q × q matrix), b_r (q × 1 vector).

$$r_t = \sigma(W_{rx}x_t) + U_{rh}h_{t-1} + b_r) \quad (15)$$

GRU reveals to the entire state of each iteration. In the same way, the candidate state $\hat{h}_t$ is similarly computed to the existing recurrent unit.

$$\hat{h}_t = tanh(W_{\hat{h}}x_t) + r_t * U_{\hat{h}}h_{t-1} + b_{\hat{h}}) \quad (16)$$

The candidate state $\hat{h}_t$ at the current timestep t, the reset gate r_t is handle the flow of the previous hidden activation h_{t-1} containing past information. If the reset gate is around zero, the previous hidden calculated state h_{t-1} will be removed.

Output state:

$$h_t = (1 - z_t) * h_{t-1} + z_t * \hat{h}_t \quad (17)$$

The hidden state h_t uses the update gate z_t to update the previous hidden state h_{t-1} and the candidate hidden state $\hat{h}_t$. If the update gate is close to 1, the previous hidden state will be held and passed to the current instant. * is the Hadamard product between of the previous state $h_{t-1}$with $(1 - z_t)$ and element-wise multiplication * of the update gate z_t with candidate activation state $\hat{h}_t$.

GRU can preserve memory substantially longer than existing RNN due to gating mechanism. However, based on the recent studies and practically observation, we find out that when GRU examines a word it only considers the forward linguistic context, so it is very needed for GRU to learn the contexts by backward pass. We also observe that the meaning of word in any language model is affected not only the forward pass but also on the backward pass.

Therefore, we proposed Two-State GRU to handle the above problem; the proposed TGRU network contain two directions, one for positive time direction namely (forward state), and other for negative time direction namely (backward state) as presented in Fig. 5. TGRU learns the contexts of a word from both directions. TGRU is inspired by the bidirectional recurrent neural networks (BRNNs) in [25]. In the training process, it splits each training sequential process into both individual recurrent networks forward and backward directions, and finally these directions are jointly combine into the output layer. The equations for update gate z_t, reset gate r_t, candidate state $\hat{h}_t$, and final output activation state h_t of the forward and backward GRU are presented as a follows:

Forward Pass:

$$\overrightarrow{z_t} = \sigma(\overrightarrow{W_{zx}x_t}) + \overrightarrow{U_{zh}h_{t-1}} + \overrightarrow{b_z}) \quad (18)$$

$$\overrightarrow{r_t} = \sigma(\overrightarrow{W_{rx}x_t}) + \overrightarrow{U_{rh}h_{t-1}} + \overrightarrow{b_r}) \quad (19)$$

$$\overrightarrow{\hat{h}_t} = \tanh(\overrightarrow{W_{\hat{h}}}x_t) + \overrightarrow{r_t} * \overrightarrow{U_{\hat{h}}h_{t-1}} + \overrightarrow{b_{\hat{h}}}) \quad (20)$$

$$\overrightarrow{h_t} = (1 - \overrightarrow{z_t}) * \overrightarrow{h_{t-1}} + \overrightarrow{z_t * \hat{h}_t}) \quad (21)$$

In addition, we added backward pass to our proposed approach to discover more useful information.

Backward Pass:

$$\overleftarrow{z_t} = \sigma(\overleftarrow{W_{zx}}x_t) + \overleftarrow{U_{zh}}\overleftarrow{h_{t-1}} + \overleftarrow{b_z}) \quad (22)$$

$$\overleftarrow{r_t} = \sigma(\overleftarrow{W_{rx}}x_t) + \overleftarrow{U_{rh}}\overleftarrow{h_{t-1}} + \overleftarrow{b_r}) \quad (23)$$

$$\overleftarrow{\hat{h}_t} = \tanh(\overleftarrow{W_{\hat{h}}}x_t) + \overleftarrow{r_t} * \overleftarrow{U_{\hat{h}}}\overleftarrow{h_{t-1}} + \overleftarrow{b_{\hat{h}}}) \quad (24)$$

$$\overleftarrow{h_t} = (1 - \overleftarrow{z_t}) * \overleftarrow{h_{t-1}} + \overleftarrow{z_t * \hat{h}_t}) \quad (25)$$

To the initiation of a word at time t: $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$ = for a random series $(x_1, x_2, \ldots, x_n)$ consisting n words, each word shown as a dimensional vector at time t.

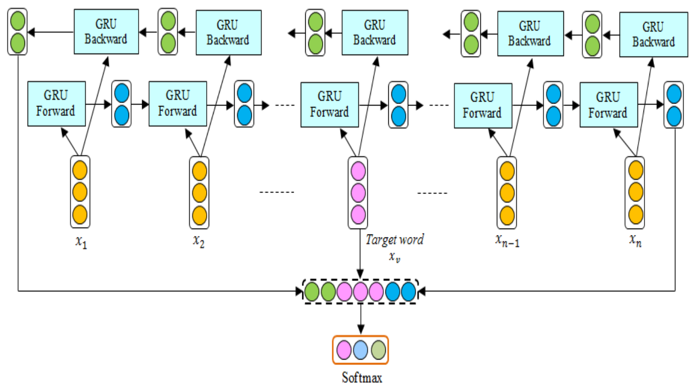


Fig. 5. The Proposed Two-State GRU Architecture for Sentiment Classification.

The forward pass of GRU calculates $\overrightarrow{h_t}$ that shows the left to the right contexts of the sentence while the backward pass of GRU phase take the right to left contexts $\overleftarrow{h_t}$ respectively. Then contexts representation of forward and backward directions were conbined into a single layer. Fig. 5 presents the detailarchitecture of TGRU.

Generally, the complexity of an model is calculated by *O(D)* and the model estimated parameters is computed by *D*. Two general pieces of information applied to calculate *D* is the dimension of the input vector p-dimension and hidden layer dimension q-dimension. Table I illustrates the detail of computed parameters of GRU, LSTM, Bi-LSTM and TGRU.

The complexity of TGRU double compare to standard GRU because the number of parameters are double. The performance of TGRU requires more time and resources for execution than GRU or LSTM while require less time and resoreces for execution than Bi-LSTM. However, our proposed E-TGRU approach is capable to explored useful information which greatly improve the accuracy of the sentiment classification.

G. Flowchart

In this sub-section we illustrate the flowchart of model for semtiment classification algorithm. Fig. 6 illustrates a flowchart of sentiment classification that contain three major phases. The 1st phase contain on data formatting for sentiment classification purpose. The 2nd phase consist of preprocessing data of model using variant GRU. In this phase, we apply the encoder to preprocess the text data. After that, the preprocessed data are utilized as input to the TGRU. The 3rd phase is cross verification.

TABLE. I. NUMBERS OF PARAMETERS

Model	Number of parameters
GRU	$3 \times (q^2 + qp + q)$
LSTM	$4 \times (q^2 + qp + q)$
Propose E-TGRU	$\boldsymbol{6 \times (q^2 + qp + q)}$
Bi-LSTM	$8 \times (q^2 + qp + q)$

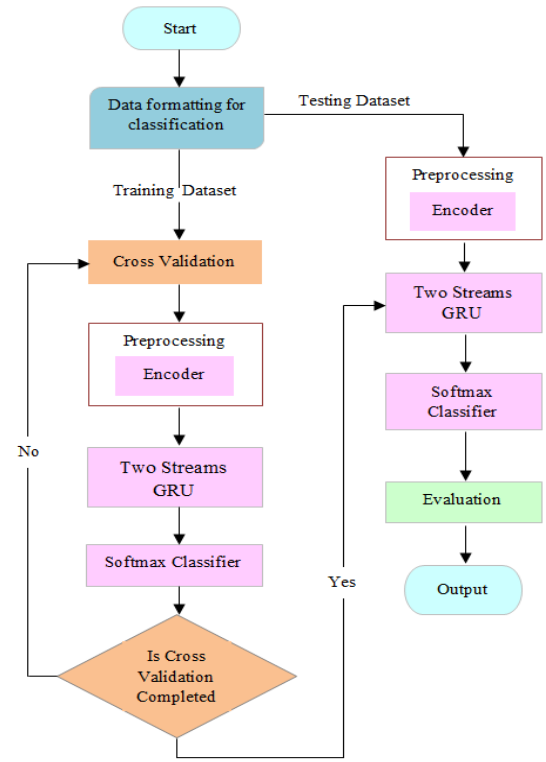


Fig. 6. Flowchart of Network Sentiment Analysis.

Algorithm 1: Encoder-two State Gated Recurrent Unit for for Sentiment Analysis

Input: vector V, that include U model features in the network, each network word as a vector $(v_{i1}, v_{i2}, \dots, v_{ir})$.
Output: Evaluates result in the dataset.

1. **Step 1: Create encoder layer**
2. Insert the 1st encoder layer of E_1 unit included *tanh* activation
3. Insert the 2nd encoder layer of E_2 unit included *tanh* activation.
4. **Step 2: Build Two-State GRU model**
5. Insert the 1st GRU layer of L_1 units with Sigmoid activation and dropout is d_1 and recurrent dropout is rd_1.
6. Insert the 2nd GRU layer of L_2 units with Sigmoid activation and dropout is d_2 and recurrent dropout is rd_2.
7. **Forward Pass:**
8. Initialize from the input layer do a forward pass over the network and take left to right contexts $\overrightarrow{h_t}$ of the sentence.
9. **Backward pass:**
10. Initialize from output layer to do backward over the network and take the right to left contexts $\overleftarrow{h_t}$ of the sentence.
11. **Step 3: Train and validate model**
12. **while** initial stop condition is not met **do**
13.   **while** training dataset is not empty **do**
14.   Prepared a mini-batch dataset as network inputs.
15.   Calculate categorical cross entropy loss function.
16.   Update weights and bias using RMSprop optimizer algorithm.
17.   **end while**
18.   Validates network with validation set.
19. **end while**
20. **Step 4: Test model**
21. Test fine-tuned hyper-parameters with test dataset.
22. **return** Evaluates result in test dataset.

It is observing the data preprocessing is very important for increasing the accuracy of the network, because the valueable features are attained by data preprocessing directly affect the final performance of the model. Therefore, we introduced variant GRU include encoder method to preprocessing data.

IV. Experimental Results and Analysis

In this section, we conducted the different experiments to present how E-TGRU performs better as compared with three well-known state-of-the-art recurrent models on two benchmark sentiment classification datasets the Stanford Large Movie Review dataset (IMDB) and Amazon Product Reviews (APR) dataset.

A. IMDB

The results for this paper were achieved using the IMDB dataset originally collected by Andrew Maas [26]. It consists of the labeled dataset of 50,000 IMDB movie reviews, especially selected for sentiment classification. These reviews are divided 50:50 ratio into training and test data. We preprocessing the dataset following the implementation similar in [27]. Furthermore, it has 50,000 un-labelled movie reviews which we useful for unsupervised training.

B. Amazon Product Reviews(APR)

We used the dataset containing Health and Personal Care product reviews from Amazon website that was available in University of California [28]. We trained the data set that contains 10,000 reviews included 50% positive and negative reviews are including to preparing it a binary classification. Additionally, these reviews are divided 15% of the dataset is applied for testing and 15% for validation aim.

C. Implementation Detail

The Data preprocessing and manipulate have performed in 3.6 Python version and anaconda. The network was trained through 30 epochs. All the simulation works were performed on Intel Core i7-3770XPU on a Windows PC with @3.40 GHz, and 4GB RAM machine. We called the name of proposed approach E-TGRU to pre-trained word vector for sentiment classification.

In this sub section, we describe each layer of E-TGRU model in detail. In order to improve the performance of the proposed model, that first step is improve the quality of the dataset, we enhance the quality of text dataset by preprocessing technique, and then gain 300-dimensional word vector by using pre-trained word2vec method selected from Google for sentiment classification. In our experiment, we used RMSprop optimizer to set their default optimal parameter setting with learning rate is 0.001 and decay factor is 0.9. The model is trained by mini-batch to performed gradient descent with batch size of 64. The sigmoid activation has applied as dense layer for binary sentiment classification and softmax activation function for multiclass sentiment classification. To avoid the overfitting problem, we have applied dropout strategy for TGRU layer with 128 memory units for each forward and backward direction. We set dropout rate of 0.2 uses by embedding layer, while the recurrent structure has a dropout of 0.4. After combining the forward and backward GRU, one more dropout layer was added to reduce 50% overfitting issue. Moreover, 10-fold cross validation has applied to minimize the arbitrary impact of the model.

Fig. 7 is presenting the detail accuracy and loss function of the model. As illustrated in the figure, after 10 epochs the training and validation accuracy greatly improve over 86% and the training and validation loss reduce below 35%. And after 25 iterations the model finally achieved accuracy over 88% and loss decrease below 30%.

D. Sentiment Analysis Results

Fig. 8 summarized the classification results on IMDB, and APR datasets. We evaluated the efficiently of our proposed E-TGRU model and compared it with three state-of-the-art exiting RNNs approaches GRU, LSTM and Bi-LSTM. The results prove that our proposed model is suitable for sentence level sentiment classification with higher accuracy of 89.37%. In our research, to train the model by using Word2vec word vector method for computing a real value vector representation of a word and performs excellent on the both IMDB and APR datasets. We fixed both the word embedding dimensions and number of units to 64.

Additionally, the outperformance of E-TGRU through encoder method and two-state GRU mechanism to demonstrate that our model is much better than other traditional models for the task of sentiment classification. Although, it can obtains the higher performance in both binary and multiclass sentiment classification tasks. In this paper, we evaluated the classification performance of the proposed models based on three evaluation metrics, namely the accuracy, F1-scores and mean square error (MSE). First, we conduct the experiment on IMDB movie reviews dataset, the performance evaluates between three baseline models are presented in Fig. 8.

In Fig. 8 shown that the proposed model GRU-Embed have achieved better performance on IMDB dataset with accuracy of 89.37%. We can see the continuously excellent performance of E-TGRU than GRU, LSTM and Bi-GRU. Next, we evaluate the classification performance of Amazon Products Reviews (APR) dataset in Fig. 9. The proposed model E-TGRU also achieved best performance on APR dataset with accuracy of 87.58%, while GRU achieved 83.08% LSTM attained 83.63% and Bi-LSTM obtained much better accuracy of 84.96% than GRU and LSTM. On the other hand, we compared F1-score performance of two traditional recurrent models is presented in Fig. 10. It is clearly showing that better performance is achieved over the application of pre-trained word embeddings. The F1-scores results show that the pre-trained word vectors are better general features extractor and can be used entire datasets.

E. Comparing Tgru With Recent Studies

This research have evaluated the networks performance compared with existing recent researches [18], [29] and [30]. Kim et al. [18] proposed static and non-static convolutional neural network trained on top of pre-trained word vector using Word2Vec for sentence level classification. They applied multiple channels have obtained 81.58% accuracy using same dataset. Socher et al. [29] introduced recursive matrix-RNN network that learns compositional vector representations for phrases and assigns the word vector and matrix method to every node in a parse tree to achieved 79.00% . Zulqarnain et al. [30] propose gated recurrent unit based on batch normalization for sentence classification. They applied batch normalization technique in forward layer and used Glove word vector in embedding layer.

Fig. 7. Effect of Accuracy and Loss at Various Time Periods.

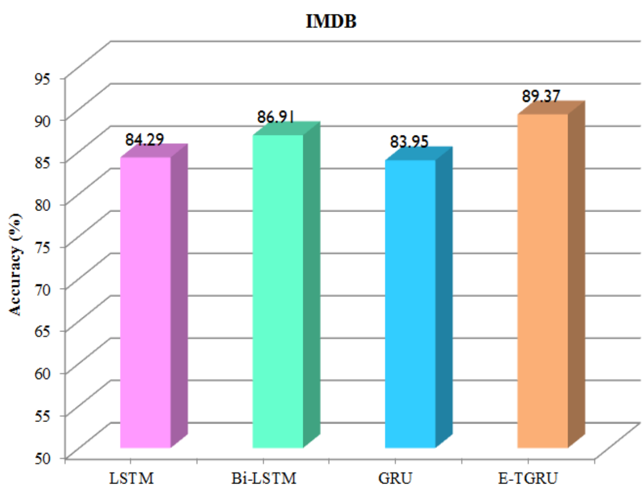


Fig. 8. Accuracy Results Comparison of Proposed E-TGRU Model with Three Baseline RNNs Models.

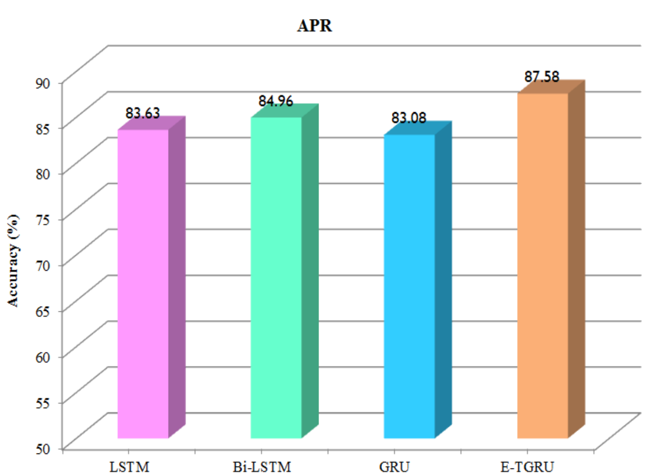


Fig. 9. Classification Performance Comparison for APR Dataset.

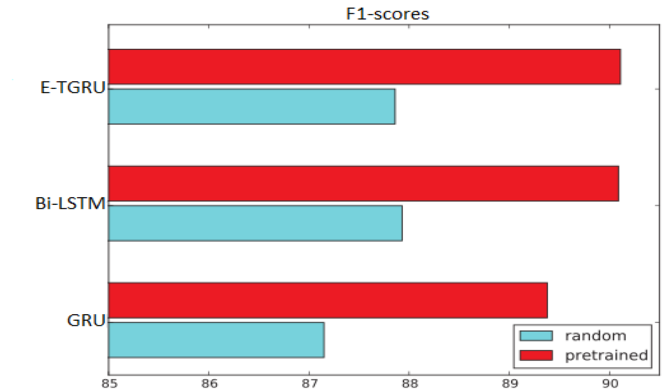


Fig. 10. E-TGRU Model Compared with Bi-LSTM and GRU in the Term of F1-Score (%) Performance by Impact of Initialize Word Embedding.

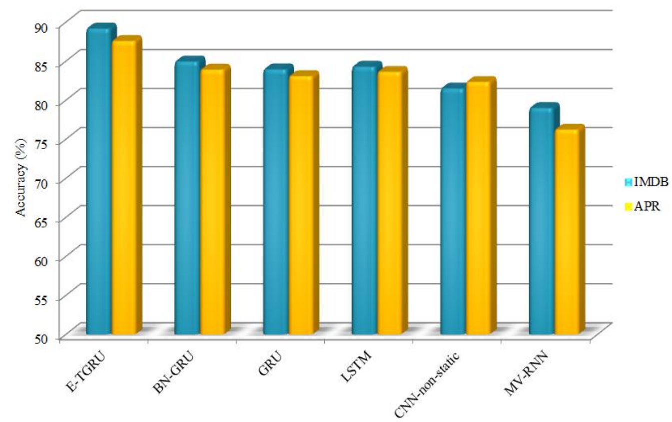


Fig. 11. Our Proposed Model Compared with Three Existing Studies in the Term of Accuracy.

LSTM, GRU and E-TGRU were executed on the same datasets and same period with three approaches. Bi-LSTM (bidirectional LSTM), which is extended variation of LSTM that has been extensively applied in recent studied [31], is also included for further comparisons. We deployed LSTM and GRU on same dataset with similar parameters to those in [32]. Fig. 11 showed the result that our proposed E-TGRU model outperforms than other existing models. This improve development is justifiable because not only we build the encoder and Word2Vec embedding method in GRU, but we also combine the forward and backward contexts to learn more useful information.

F. Comparison Error Rate with Traditional RNNs

In this section, we perform to analysis an error rate of our proposed E-TGRU model with three state-to-the-art RNNs approaches such as standard GRU, LSTM and recent Bi-LSTM. Execution setup showed that with the continuous increase of epochs, the mean square error is continuously decreasing and the final MSE is 0.0162 on IMDB dataset and 0.2713 on APR. We fixed both the word embedding dimensions and number of units to 64 and execute the model for 30 epochs. We found that proposed model converged faster than GRU, LSTM and Bi-LSTM to achieved very lower error rate even after many epochs. To make these models comparable, we implement these models with the identical structural design. Finally, we evaluate our E-TGRU model with state-of-the-art existing RNNs models on IMDB and APR datasets. Table II demonstrates the results that proposed E-TGRU model achieves much better performance in the term of the error rate than GRU, LSTM and Bi-LSTM.

TABLE. II. COMPARISON ERROR RETE (%) WITH EXISTING RNNS MODELS

Models	**IMDB**	**APR**
LSTM	*0.1278*	*0.3569*
GRU	*0.1025*	*0.3490*
Bi-LSTM	*0.0842*	*0.3125*
E-TGRU	***0.0162***	***0.2713***

V. CONCLUSION

Sentiment classification remains popular and significant area of natural language processing. In this paper, we investigated variant gated recurrent unit included encoder GRU (E-GRU) to preprocess the texts data for sentiment classification. E-GRU frequently provides an excellent representation of the input than the original raw input. Furthermore, we also developed Two-State Gated Recurrent Unit (TGRU) which is included forward and backward states, that is capable to learn more valuable information, especially for text processing issue. Then, we used Word2Vec pre-trained word embeddings method that is possible to learn the contextual semantics of words from the text can be effectively classified. Based on experimental observation, we found that RNNs models, being a recurrent network it can effectively capturing the useful information from a massive array of sequential data and the best choice in terms of accuracy. We conduct the experiment on two benchmark sentiment analysis datasets, included IMDB and APR respectively. The proposed E-TGRU model achieved highest 89.37% accuracy on IMDB dataset and 87.58% accuracy on APR dataset. Our proposed model achieves much better performance in the term of an error rate than GRU, LSTM and Bi-LSTM, when increases the number of epochs.

In future work, there are many ways to extend this work. Future research can be dedicated the proposed approach using multiple sentiment lexicons and much powerful ranking approaches to enhance the sentiment classification performance and also reduce the computational complexity of the proposed model.

ACKNOWLEDGMENT

The authors would like to thanks Ministry of Education Malaysia and Universiti Tun Hussein Onn Malaysia (UTHM) for funding this research activity under the Fundamental Research Grant Scheme (FRGS), Vot Number: FRGS/1/2019/ICT05/UTHM/03/1.

REFERENCES

[1] P. Vincent, "A Neural Probabilistic Language Model," A neural probabilistic Lang. Model. J. Mach. Learn. Res., vol. 3, pp. 1137–1155, 2003.

[2] M. Baroni, "Composition in distributional semantics," Lang. Linguist. Compass, vol. 07, no. June, pp. 511–522, 2014.

[3] P. D. Turney and P. Pantel, "From Frequency to Meaning : Vector Space Models of Semantics," J. Artif. Intell. Res., vol. 37, pp. 141–188, 2010.

[4] M. Aamir, F. Wahid, H. Mahdin, and N. M. Nawi, "An efficient normalized restricted Boltzmann machine for solving multiclass classification problems," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, pp. 416–426, 2019.

[5] C. Burgess, C. Burgess, K. Livesay, and K. Lund, "Explorations in Context Space : Words , Sentences , Discourse," Discourse Process., vol. 25, no. June, pp. 211–257, 2015.

[6] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient Estimation ofWord Representations in Vector Space," arXiv Prepr. arXiv1301.3781, pp. 1–12, 2013.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "5021-Distributed-Representations-of-Words-and-Phrases-and-Their-Compositionality," Adv. Neural Inf. Process. Syst., vol. April, pp. 3111–3119, 2013.

[8] M. Zulqarnain, R. Ghazali, M. G. Ghouse, and M. F. Mushtaq, "Efficient Processing of GRU Based on Word Embedding for Text Classification," Int. J. Informatics Vis., vol. 3, no. 4, pp. 377–383, 2019.

[9] L. Ratinov and J. Turian, "Word representations : A simple and general method for semi-supervised learning," Proc. 48th Annu. Meet. Assoc. Comput. Linguist., no. July, pp. 384–394, 2010.

[10] K. Gimpel, "Tailoring Continuous Word Representations for Dependency Parsing," Proc. 52nd Annu. Meet. Assoc. Comput. Linguist., vol. 2, pp. 809–815, 2014.

[11] W. Sharif, N. A. Samsudin, M. M. Deris, and M. Aamir, "Improved relative discriminative criterion feature ranking technique for text classification," Int. J. Artif. Intell., vol. 15, no. 2, pp. 61–78, 2017.

[12] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light Gated Recurrent Units for Speech Recognition," IEEE Trans. Emerg. Top. Comput. Intell., vol. 2, no. 2, pp. 92–102, 2018.

[13] S. M. Mohammad, B. J. Dorr, G. Hirst, and P. D. Turney, "Computing Lexical Contrast," Comput. Linguist., vol. 39, no. January 2010, pp. 555–590, 2013.

[14] D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou, "Sentiment Embeddings with Applications to Sentiment Analysis," IEEE Trans. Knowl. Data Eng., vol. 28, no. October, pp. 496–509, 2016.

[15] S. Clark, "Vector Space Models of Lexical Meaning," Handb. Contemp. Semant. Ed., no. September, pp. 1–42, 2012.

[16] J. Pennington, R. Socher, and C. D. Manning, "GloVe : Global Vectors for Word Representation," Proc. Conf. Empir. Methods Nat. Lang. Process., no. October, pp. 1532–1543, 2014.

[17] A. Aizawa, "An information-theoretic perspective of tf – idf measures q," Inf. Process. Manag., vol. 39, pp. 45–65, 2003.

[18] Y. Kim, "Convolutional Neural Networks for Sentence Classification," Artif. Intell. Rev., no. 4, pp. 655–665, 2014.

[19] P. Liu, X. Qiu, and X. Huang, "Recurrent Neural Network for Text Classification with Multi-Task Learning," Proc. 25th Int. Jt. Conf. Artif. Intell. IJCAI-16, p. to appear, 2016.

[20] K. Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," arXiv, no. September, pp. 1–15, 2014.

[21] S. Hochreiter, "Long Short Term Memory," Neural Comput., vol. 9, no. 8, pp. 1–32, 1997.

[22] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," Neurocomputing, vol. 184, no. November, pp. 232–242, 2016.

[23] A. H. Ombabi, O. Lazzez, W. Ouarda, and A. M. Alimi, "Deep Learning Framework based on Word2Vec and CNN for Users Interests Classification," 2017.

[24] K. Cho, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," arXiv, vol. 5, pp. 1–9, 2014.

[25] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, 1997.

[26] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," Proc. 49th Annu.

Meet. Assoc. Comput. Linguist. Hum. Lang. Technol., vol. 1, pp. 142–150, 2011.

[27] P. Lamblin et al., "Theano : new features and speed improvements," arXiv Prepr. arXiv1211.5590, pp. 1–10, 2012.

[28] J. Mcauley, R. Pandey, and J. Leskovec, "Inferring Networks of Substitutable and Complementary Products," Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. data Min., pp. 785–794, 2015.

[29] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic Compositionality through Recursive Matrix-Vector Spaces," Proc. 2012 Jt. Conf. Empir. methods Nat. Lang. Process. Comput. Nat. Lang. Learn., no. July, pp. 1201–1211, 2012.

[30] M. Zulqarnain, R. Ghazali, S. H. Khaleefah, and A. Rehan, "An Improved the Performance of GRU Model based on Batch Normalization for Sentence Classification," Int. J. Comput. Sci. Netw. Secur., vol. 19, no. 9, pp. 176–186, 2019.

[31] P. Zhou et al., "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification," Proc. 54th Annu. Meet. Assoc. Comput. Linguist. (Volume 2 Short Pap., pp. 207–212, 2016.

[32] H. Peng, E. Cambria, and A. Hussain, "A Review of Sentiment Analysis Research in Chinese Language," Cognit. Comput., vol. 9, no. 4, pp. 423–435, 2017.

Angle Adjustment for Vertical and Diagonal Communication in underwater Sensor Networks

Aimen Anum[1], Tariq Ali[2]
Department of Computer Science
COMSATS University Islamabad
Sahiwal Campus, Sahiwal
Pakistan

Shuja Akbar[3], Iqra Obaid[4], Muhammad Junaid Anjum[5], Umar Draz[6], Momina Shaheen[7]
Department of Computer Science
COMSATS University Islamabad, Lahore Campus
Lahore, Pakistan

***Abstract*—Underwater wireless sensor network has been an area of interest for a few previous decades. UWSNs consists of tiny sensors responsible for monitoring different underwater events and transmit the collected data to the sink node. In the harsh and continuously changing environment of water, gaining better communication and performance is a difficult task as compared to networks available on land because of different underwater characteristics such as end-to-end delays, node movement, and energy constraints. In this paper, a novel routing technique named angle adjustment for vertical and diagonal communication was proposed, which doesn't use any location information of nodes. It is also efficient in terms of energy and end-to-end delays. In this approach, the source node evaluates the flooding zone based on the angle by using the basic formula for forwarding the packet to the sink. After evaluating the flooding zone, the angles of each node are compared and the packet is sent to the node closest to the vertical line. The proposed approach is evaluated with the help of NS-2 with AquaSim. The results show better results performance in data delivery, end-to-end delays, and energy consumption than DBR.**

Keywords—Wireless Sensor Networks; Underwater wireless sensor network; DVRP; Terrestrial Wireless Sensor Networks; Depth Based Routing (DBR)

I. Introduction

Wireless Sensor Network (WSN) is a worldwide interest. It is a network that consists of tiny nodes that are equipped with power supply, analog to digital converter, small memory, processor and a radio interface. Sensors have the capability to communicate with other sensors. WSN is divided into different groups such as Terrestrial Wireless Sensor Network (TWSN) and Underwater Sensor Network (UWSN) etc. In TWSN, nodes are deployed on land and must be able to communicate effectively. Nodes can have a secondary power source. Whereas the architecture of UWSN has sensors randomly deployed in the form of one dimension, two dimensions, three dimensions, and four dimensions in the oceanic environment flood data forwarding towards the base station and have limited battery and are also expensive.

Ocean covers more than 70% of the earth. UWSN is gaining lots of attention because oceans are playing vital role in transportation, ocean exploration, defense, obtaining valuable minerals, oil and gas, surveillance data collection and adventurous means [1][2][3]. It also helps in avoidance of different disasters i.e. flood, tsunami, submarine detection and pollution control [4][5]. Underwater Exploration: Sensors helps in extracting resources such as rare metals and minerals, monitoring of equipment, oil platforms, buried communication cables, and gas pipes. Sensors are also used in the localization of objects and in the successful discovery of many lost treasures. Seismic Monitoring: Water resources hides most of the oil and gas resources. To extract the oil and gas resources, frequent 3D, and 4D seismic monitoring is required [3].

Mine Reconnaissance: Early detection of minefields to avoid disasters was always difficult. UWSNs can effectively help as a tool for early detection and deactivation.

Disaster Prevention: In UWSNs, research is conducting in the area of disaster prevention and recovery [2]. With the help of UWSNs, early warnings of the tsunami can be provided to coastal. It can also aid in oil spills, disaster recoveries, and investigation of marine incidents.

Assisted Navigation: Sensors can be helpful in locating different harmful objects such as rocks, shoals, and wrecks in shallow as well as deep water.

Underwater Sensor Network and terrestrial wireless networks have some resemblances but are not applicable in underwater environments as radio signals cannot propagate in water, but sound signals are responsible for propagation [6]. So the terrestrial protocols cannot be used in an underwater environment [7]. Routing in the underwater environment is challenging to make its applications reliable [11] and also because of its multiple limitations which are more power consumption, high error rates, limited bandwidth, high end-to-end delay, no recharging mechanism, and continuous node movements [12][8][13]. For the efficient use of energy, multi-hop communication is preferred [12]. UWSN protocols are classified into two classes: 1) Localization Based Routing Protocols, 2) Localization Free Routing Protocols. Localization based protocols require accurate information of localization for packet forwarding which is the main liability. Localization free protocols can depend upon different parameters such as link strength, residual energy, and depth. A localization free routing protocol named Depth Based Routing Protocol (DBR) was proposed which does not require any geographical information but only needs the information about the depth of each node. The source node broadcasts the data to all its neighbors having a depth less than the source node. The process continues until the data is received at the sink [8] [9]

[10]. But broadcasting the data can degrade network performance [11]. To avoid horizontal communication a Diagonal and Vertical Routing Protocol (DVRP) was proposed which uses angle-based flooding. In this protocol, the nodes flood the data either vertically or diagonally because the vertical or diagonal distance is always less than the horizontal distance towards the sink. Nodes flood the data to nodes at the upper layer with the help of a formula $\theta=90\pm10K$. The flooding zone is always greater than 0 and less than π [7] [12].

A. Challenges of underwater Wireless Networks

- Limited Energy: Harvesting and consumption of energy is one of the key challenges in UWSN. Underwater sensors are of large size; hence they need more energy for communication [5]. Sensors are deployed in the deep sea where sunlight is unavailable which means recharging of batteries is very difficult. UWSN must focus on power efficient designs and some power-efficient and robust protocols should be developed [13].

- Low Data Rates: Although acoustic communication is ideal in an underwater environment for its long transmission range and reliability. But acoustic communication offers the data of 5kbps to 20kbps which is extremely low as compared to radio communication which offers data rate in Gbps.

- Propagation Delays: The acoustic communication provides a speed of 1500m/sec which is five times less than that of radio communication [5]. The slow speed produces a delay of 0.67s/km [14]. These delays decrease the throughput of the network.

- Dynamic Topology: Dynamic topology is one of the challenges of the underwater environment. As the nodes move 1-3m/sec with the water currents [16]. Due to the movement of nodes, it is difficult to maintain a static topology which greatly effects the performance of routing protocols [5][17]. To handle such a situation, the routing protocol must have the information of nodes' location or depth [5].

- Reliability: The oceanic environment is very unpredictable for communication because of Doppler spread, pressure, marine life, man-made noise, salinity, and ocean currents. In the water, nodes move about 2-3m/sec due to water currents which effects in localization of nodes, communication link, and the network topology. All of these results in unreliable data transmission [15].

- Noise: Noise is the communication quality due to which strength of signal degrades. Noise can be produced by fishing, ships, human, and marine activities [18]. Noise is classified into two classes which are man-made noise and ambient noise.

    - Man-Made Noise: The noise which is produced by different activities such as fishing, use of machines, sonar, shipping, and military [18]. These activities produce interference and disturbance in communication [19].
    - Ambient Noise: It is a complex process that occurs by a combination of different undefined sources which cannot be identified uniquely [20]. It is also known as background noise. Ambient noise is majorly produced by factors like wind, shipping, and turbulence [21]. The destruction of waves due to air bubbles is referred to as wind noise. Ships can be considered as a major cause of ambient noise. The presence of ships at a large distance from communication can produce high noise ratios. The tides produce a disturbance in the surface of the water. This disturbance results in continuous low-frequency noise [19].
    - Multipath: Horizontal communication channel is highly affected by multipath as compared to vertical channel. Reflections weaken acoustic signals, produce long delays, inter-symbol interference and make the data erroneous [22][23][24] [25].
    - Doppler Spread: The shift of mean frequency due to relative motion of transmitter and receiver is known as the Doppler shift. Frequency fluctuation in the area of the Doppler shift is termed as doppler spread [19][26].

B. Research Objectives

The key challenge in UWSN is to receive data packets on the surface of the water. The proposed algorithms for vertical and diagonal communication are based on an angular technique that causes the dark shaded areas in communication. The nodes inside the dark area cannot participate in the communication process due to improper adjustment of angle to restrict horizontal communication. These angle adjustments can cause for high end to end delay, long data routing path, data packet losses, and network lifetime.

A specific objective of the research is to design and develop a new Angle Optimization for Vertical and Diagonal Communication in Underwater Sensor Networks to:

1) Minimize the dark shade area.
2) Optimal adjustment of angle
3) Reduce the horizontal communication among nodes.
4) Overcome high end to end delays.
5) Increase throughput and network life-time.

These Objectives are transformed into the form of research questions.

RQ1: How dark shaded area increases end-to-end delay?

RQ2: How can we design the protocol to adjust the angle for the optimal solution of the dark shaded area?

RQ3: How to decrease the delay caused by the dark shaded area?

RQ4: How the proposed algorithm can increase the throughput for efficient communication?

II. Related Work

About 70% of the surface of the earth is covered with water which is in the form of rivers, canals, and oceans. Oceans have a huge amount to precious resources such as oil and gas. The exploration of these resources depends on the technology. Recent advancements in technology help in the exploration of resources also for the avoidance of disasters, detection of pollution and data collection. Technology that is applicable to all these applications is UWSN [1]. The architecture of UWSN consists of three parts which are sensors, sinks and surface stations. The data is sensed by sensor nodes and forwarded towards the surface station by using sinks [27][25]. The architecture of UWSN has sensors randomly deployed in the form of one dimension, two dimensions, three dimensions, and four dimensions.

A. *One Dimension UWSNs*

1D architecture uses the autonomous deployment of sensors. Every node is a singleton network which performs the task of sensing, processing, and transmission of information towards remote station. Each node is a floating buoy having sensing capabilities deployed for a specified interval of time to gather information. After this specified time, it floats back towards the surface to send data to the remote station. Nodes can also be an AUV that dives into the water to sense underwater activities and carry the information to the remote station. 1D can use acoustic, optical or radio waves as the communication medium. The communication can be transmitted to remote station only by using single-hop [15].

B. *Two Dimension UWSNs*

In 2D underwater sensor networks, several nodes are deployed in the seabed with anchors nodes. Sensors are connected to the sink node via acoustic signals as shown in Fig. 1. Sink nodes are responsible for the transmission of data from sea bed to surface. Each underwater sink has a horizontal and vertical transceiver. To perform data collection and transmission of configuration commands between sensor nodes and sink nodes, it uses horizontal communication. Whereas vertical transceiver is used to send data to surface station. The surface station has an acoustic transceiver to communicate with the sink node and radio transceiver for communication with the surface sink [28]. For connecting with underwater sinks, sensors nodes can either use direct link or a multi-hop route. Using a direct link is the simplest method for sending data, but it is not energy efficient. The use of direct links results in a decrease in network throughput and an increase in interference. On the other hand, in multi-hop communication, data collected from the sea bed is forwarded using intermediate sensors until it reaches the sink. This enhances network capacity, saves energy and also increases the complexity of network [24]. Terrestrial wireless sensor networks cannot be used in oceans because radio waves cannot travel in water, and the ocean's diverse environment creates many challenges such as high delays, limited bandwidth, limited battery and node mobility [19].

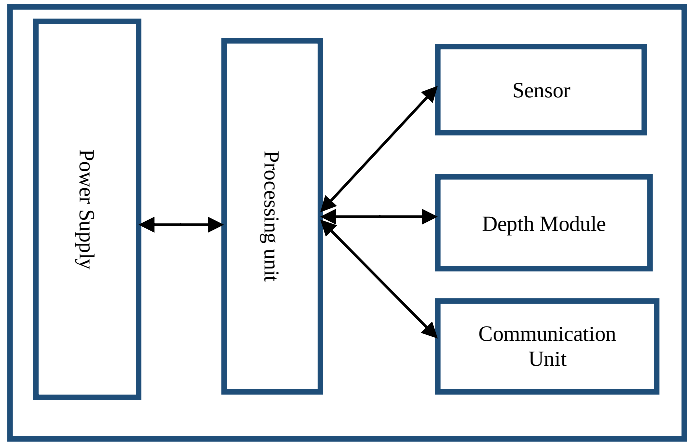


Fig. 1. Internal Architecture of underwater Node.

C. *Three Dimension UWSNs*

The 3D UWSNs are helpful in observing different processes or activities in the ocean bed. In the architecture of 3D UWSNs, sensor nodes are deployed in an ocean bed. Each node has a floating buoy with a pump. The sensors are pushed on the surface with the help of a buoy. A wire connecting sensor and anchor helps in the adjustment of the sensor's depth. The wire is adjusted by an electronic engine that is placed on sensors. The issue of ocean current was addressed in the architecture, but many other issues arise in with this architecture which are

1) *Sensing coverage:* All the sensor should adjust their depths to gain complete 3D coverage of ocean with respect to sensing ranges.

2) *Communication coverage:* Sensing nodes should have the ability to carry information towards the surface by using multi-hop links. Nodes should regulate their depths so that the topology remains connected and there must exist at least one path from each sensor to surface station [28] [24].

D. *Four Dimension UWSNs*

This architecture is a fusion of mobile and fixed networks, where fixed networks refer to 3D UWSNs and mobile network consists of remotely operated underwater vehicles (ROVs). ROVs can be submarines, ships or submersible robots. ROVs gathers the data from deployed nodes and carry it to the remote station. The nodes having a lot of data and are close to ROV can use radio waves to send data directly to ROV [15].

E. *Available transmission medium for UWSN*

Communication by using acoustic signals is widely used in underwater communications. In the conducting nature of seawater, acoustic signals propagate well in low frequencies even at long distances [29][30][31]. Although acoustic communication is affected by noise, temperature, multipath propagation. Because of the slow speed of 1500m/s acoustic medium has low bandwidth of less than 100KHz [32]. But acoustic communication is still favorable due to low attenuation rate and long transmission range of 50m to 5km as shown in Table I.

TABLE. I. COMPARISON OF AVAILABLE TRANSMISSION MEDIUM

	ACOM	EMCOMM	OCOMM
Range of Transmission	~50m-5km	~1m-100m	~1m-100m [2]
Data Rate	100Kbps[25]	10Mbps	1GBps [25]
Complexity of Antenna	Medium	High	Medium
Antenna Size	~ 0.1 m	~ 0.5 m	~ 0.1 m [20]
Power Loss	> 0.1 dB/m/Hz	~28 dB/100mHZ	∞ turbidity [20]

The most useful property of electromagnetic waves is that it uses higher bandwidth. Due to the conducting nature of seawater, radio waves cannot work properly [25]. If electromagnetic waves work in water, it will provide high data rates of 10mbps by using a highly complex antenna. But electromagnetic waves are highly affected by signal attenuation and electromagnetic interference [33][29].

The nature of light is the key reason for limited optical signals. The optical wave communication provides the fastest data rates as compared to radio and acoustic signals [34]. It offers data rates of 1Gbps. The optical signals are absorbed in water due to which its intensity reduces, attenuated, noisy and scattered in water [35][36]. As the presence of sunlight is one of the major causes of ambient noise. So, the strength of the signal will be less than the noise produced by ambient light. To reduce ambient noise optical modems, use high-pass filter technique but the use of filters increases the cost [29] [37].

Many underwater routing protocols have been introduced based on different parameters i.e. reliability, mobility, delivery ratio, and energy efficiency, etc. [38]. Some of the routing protocols forward the data on the basis of angles, two of which are: 1) Layer by Layer Angle Based Flooding and 2) Diagonal and Vertical Routing Protocol.

F. Layer Angle Based Flooding (L2-ABF)

Layer by Layer Angle Based Flooding was proposed by the author in [39]. The goal of this protocol was to avoid horizontal communication between the nodes at the same levels. The sensors do not require any localization information. For packet forwarding sensor nodes sends the packets to nodes at upper layer for investigation of eligible nodes within specified area and send it to nodes having low Hop ID at the upper layers with help of basic formula of $\theta = 90 \pm 10K$ where K has the values of 1 to 8 and the θ always lies between 0 and π. With the increment of value of K flooding zone also increases. The restriction of angles helps in the avoidance of horizontal communication [14] [21] [40].

Furthermore, in our previous work like [41-45] in which, we performed simulation under different traffic agents such that TCP and UDP. In addition, link failure detection between two nodes, subnet based approach and some relevant article to support our words, for this, see [46-50].

G. Diagonal and Vertical Routing Protocol (DVRP)

Diagonal and vertical routing protocol is an angle-based flooding protocol. In the protocol, the source node does not require any localization information for packet forwarding. The aim of developing is to reduce horizontal communication among nodes as the horizontal distance is always greater than the diagonal or vertical distance towards the sink. The reduction in horizontal communication increases network lifetime and reduces delays. When a node has a data packet to send, it sends a hello packet to upper layer nodes to get the information about nodes that are eligible for data forwarding. Packets are only sent to the eligible nodes having a depth less than the source node, and multiple nodes can be eligible for transmission at once. The specified area to which packets are forwarded is called flooding zone which is calculated by of $\theta = 90 \pm 10K$ where K has the values of 1 to 8, and the θ always lies between 0 and π. If the source node does not get a hello reply in the response of the hello packet, the angle is incremented in order to increase the flooding zone of the source node [7][18].

III. RESEARCH METHODOLOGY

The deployment of sensor nodes within the water for the purpose of communication has gained worldwide interest in the past few decades. UWSNs have many similarities with TWSNs. But, UWSNs and TWSNs are also different in many aspects. The environment and applications of UWSNs are different. Due to these differences, terrestrial routing protocols cannot be used in the oceanic environment. These challenges require optimal solutions for situations i.e. water currents, long delays, and horizontal communication. A routing scheme is required which can minimize the mentioned challenges as much as possible.

Routing is the basic need of any network. The routing protocols are responsible for the discovery and maintenance of routing paths. A lot of research has been conducted on the physical layer, whereas the network layer issues are relatively new.

A. Justification for Vertical and Diagonal Communication

In the water, the goal of communication is to transmit the data to the sink via nodes of the upper layer. The data can be sent to other nodes in the horizontal, diagonal or vertical direction. The proposed routing techniques restrict the horizontal communication and prefer either diagonal or vertical communication because the distance that is covered in either diagonal or vertical direction is always less than that of the horizontal direction.

In Fig. 2, O, A, and B are ordinary sensor nodes floating at different depths of water, whereas S is the sink node that is placed on the surface of the water. The nodes O and A are deployed at the same depth level from sink S. If a node O has a data packet to send, it can send the data via any of three routes which can be

1) O to S
2) O to B and B to S
3) O to A, A to B, A to S

For the distance comparison, only routes 2 and 3 are considered. By using route 2, O can send packet to the diagonal node B which is closer to sink and then the node B will forward the packet to sink node S. whereas, in the route 3, packet will be sent to node A which is at the same depth level of O, A will send the data to node B of upper layer and finally, B will send data to S. By the theorem of triangular inequality, we prove that vertical and diagonal distance is always less than horizontal.

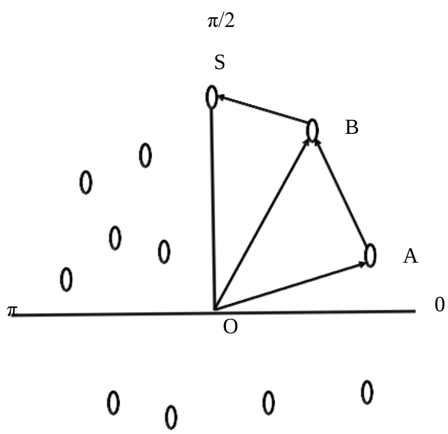


Fig. 2. Comparison of Horizontal Distance with Diagonal or Vertical Distance.

Consider the triangle ΔOAB.

$$|OB| < |OA| + |AB| \quad (1)$$

The above equation shows that the third distance is always less than the sum of two distances. So, if the packet is transmitted by the route consisting of |OA|+ |AB| it will cover more distance as compared to |OB|. Even if the addition of |BS| is done on each side of the above equation.

$$|BS| + |OB| < |OA| + |AB| + |BS| \quad (2)$$

Still, it is clear that the diagonal or vertical communication is better than horizontal communication.

B. Angle Adjustment for Vertical and Diagonal Communication

An angle adjustment technique is proposed in this research. This is a layered delay minimizing approach. It uses only diagonal and vertical communication between nodes. It uses multi-sink architecture, which gathers data from sensor nodes. Sensor nodes are deployed in multiple layers at different depths. Nodes flood the collected data within the computed angle. But when the data is flooded in the flooding zone, there may be some nodes that cannot participate in the communication creating a dark shaded area, as shown in Fig. 3. The data is sent to the upper layers until it reaches any sink. The reception of data at any sink node will be considered as a successful delivery. Sinks can transmit the data to the base station via radio signals.

The major advantages of the proposed approach are:

1) No need for localization information
2) Easily survive with node mobility occurred due to water currents
3) No maintenance of complex routing tables

The goals behind the development of this approach are as follows.

1) Minimize the dark shaded area.
2) Reduce horizontal communication among nodes.
3) Reduce propagation delays.
4) Increase the packet delivery ratio.

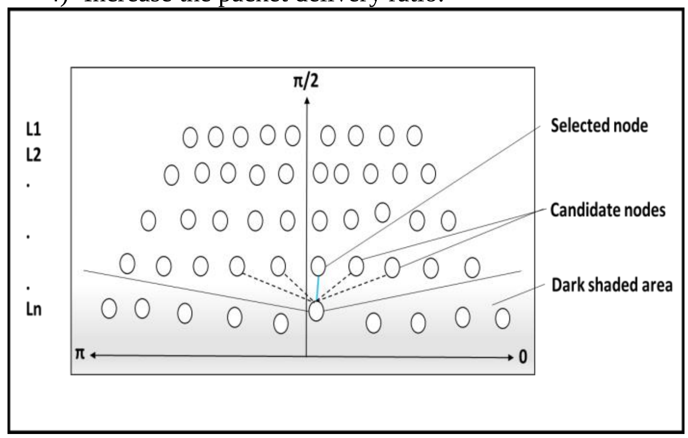


Fig. 3. Working of the Proposed Solution.

C. Control Packets

The sensor node generates a hello packet that is flooded in the flooding zone to identify its neighbors. The hello packet includes the following fields as depicted in Table II S_ID: the source ID identifies the id of the sender node.

The format of hello reply is shown in Table III. S_ID identifies the source node, R_ID identifies the ID of reply node.

1) *Data Packet Format:* The format of the data packet is shown in Table IV. The header of the data packet comprises three fields which include S_ID, R_ID, and SN_ID.

S_ID is used for the identification of the source node. F_ID consists of the id of the receiver node which was selected as a forwarder node by the sender. SN_ID is used to distinguish packets.

D. Layer ID Assigning Algorithm

For the purpose of assigning Layer IDs (L_ID) surface sink broadcasts a hello packet. At the beginning, all the ordinary nodes are assigned with L_ID 00 and the Total Layers is 9. The Total Layer is dependent on the depth of area of network. When a node receives a hello packet, it first checks its type. If the type of packet is S_hp and Total_Layer is less than i and if L_ID is equal to 00, then the value of i is assigned to L_ID. Else if L_ID is less or equal to i then the packet is discarded. If the type of packet is S_hp and Total_Layer is less than i but L_ID is not equivalent to Total_Layer then the current value of Total_Layer is assigned to L_ID.

Former this procedure, for updating L_ID the current node will send HP to the nodes of layers below it. But before sending it to lower layers the Total_Layers is decremented by 1. After decrement, if Total_Layer is equivalent to 0 then further broadcasts are not required and S_hp are discarded. The pseudocode of this Algorithm is shown in Table V and the flow is represented in Fig. 4.

TABLE. II. HELLO PACKET FORMAT

S_ID	L_ID	Angle

TABLE. III. HELLO REPLY FORMAT

S_ID	**R_ID**	**Angle**

TABLE. IV. DATA PACKET FORMAT

S_ID	F_ID	SN_ID	Data

TABLE. V. PSEUDOCODE FOR LAYER ID ASSIGNING

```
L_ID = 00       //when layer id is not assigned to node
Total_Layers=9
i=1                 // initialization of number of layers
if (P_type == S_hp) && (i < Total_Layers)
        if (L_ID = = 00)
                L_ID = i            // Each node will get layer id
else
                if (L_ID ≤ i)
                        Discard DP            // layer id is already assigned
                Else
                 L_ID = i
                    i++
                 Send S_hp further
                End if
          End if
End if
Total_Layers=9
No further broadcast for S_hp
Exit
```

1) *Updating Layer IDs:* The presented approach is suitable for both types of applications time-critical and non-time critical. The sensor nodes always remain active for time-critical applications due to which the energy usage increases. Whereas for non-time critical applications, nodes can sense and send data in a time interval after which nodes operate in sleeping mode or their transceivers are turned off.

Nodes in the water can easily move horizontally and may also slightly in a vertical direction. The movement can also change the neighbor of nodes. The Layer ID defines how many layers a packet has to cross to reach to sink node. The layer ID of the current node is always smaller than that of the lower layers and is larger than the upper layer.

In the proposed approach, after a particular time, new L_IDs are assigned to each node. When the lifetime increases a threshold value of 30 min, the entire network is assigned with new L_IDs by using HPs. If a node has a data packet ready to be sent prior to the change of L_ID, it will hold the packet till new L_ID is assigned. The procedure of new L_ID assigning can effectively tackle vertical node movement as the nodes will have new L_IDs as per their new layers.

E. Flooding Zone Calculation

The proposed approach is an angle-based flooding approach. In this approach first of all the flooding zone is computed. The flooding zone is the area of the network consisting of nodes of upper layers in which the data packets are sent. Here we made an assumption that each node knows about its *hardware* built-in base angle π/2 in the upward direction. Each node has the ability to calculate the angle and also to increase the size of its flooding zone as per requirement. The computation of flooding zone is done by using the basic formula $\theta = \pi/2 \pm \alpha$ where α is a variable having a finite set of values between 0 and $\pi/2$. The conceptual illustration of flooding zone increment is given below in Fig. 5, 6 and 7.

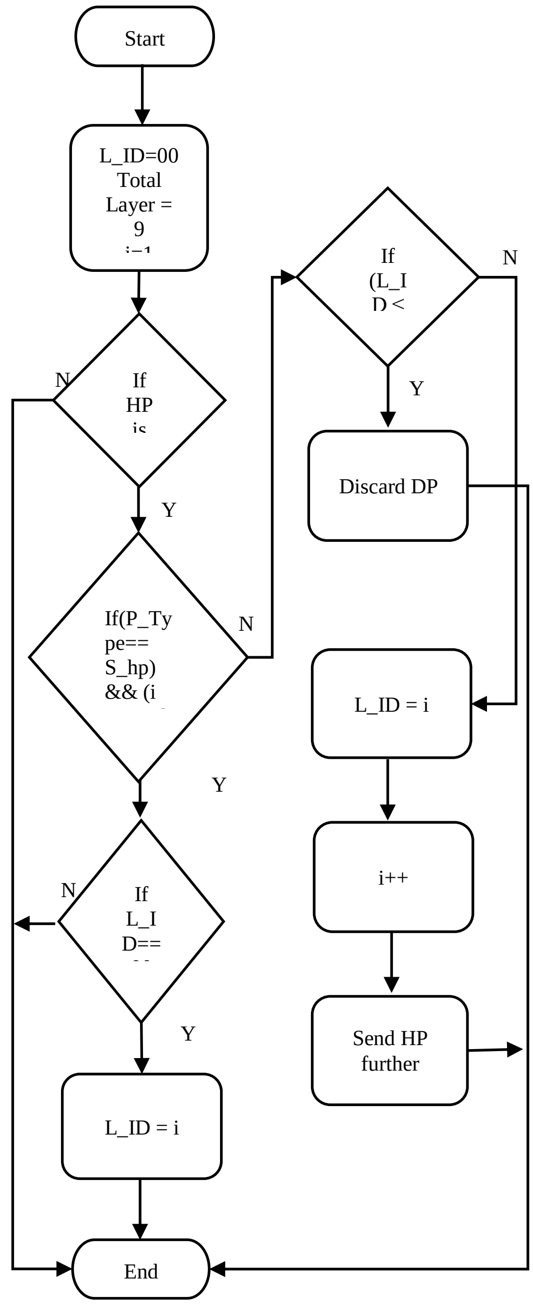


Fig. 4. Flowchart for Assigning Layer-ID.

$$\theta = \pi/2 \pm \alpha, \theta = \pi/2 \pm 1, when\ \alpha = 1 \quad (3)$$

$$\theta = 91, \Theta = -89 \quad (4)$$

$$\theta = \pi/2 \pm \alpha, \theta = \pi/2 \pm 12, when\ \alpha = 12 \quad (5)$$

$$\theta = 102, \Theta = -78 \quad (6)$$

$$\theta = \pi/2 \pm \alpha, \theta = \pi/2 \pm 89, when\ \alpha = 89 \quad (7)$$

$$\theta = 179, \Theta = -1 \quad (8)$$

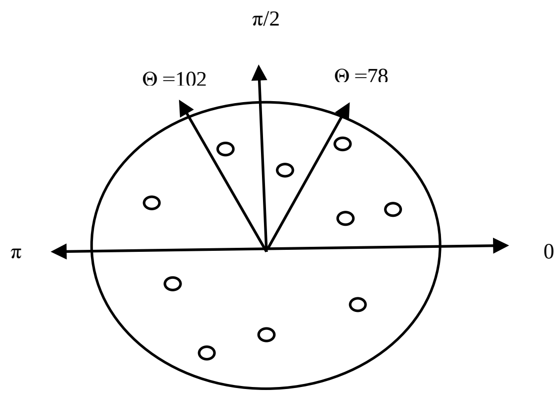


Fig. 5. Flooding Zone Calculation at α = 1.

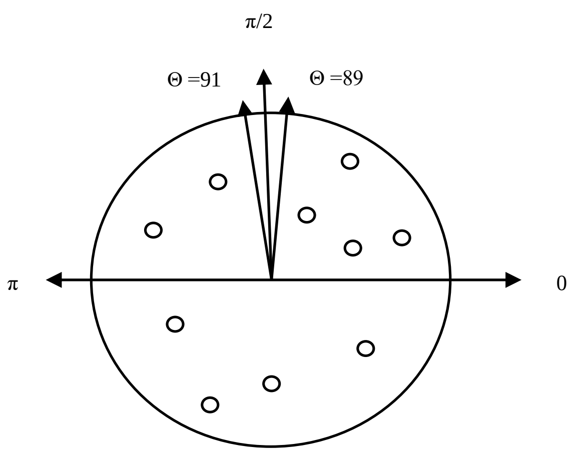


Fig. 6. Flooding Zone Calculation at α = 12.

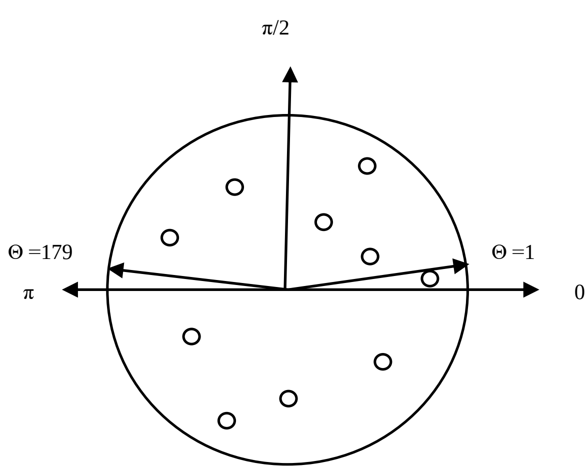


Fig. 7. Flooding Zone Calculation at α = 89.

F. Packet Transmission

The angle adjustment technique is a localization free approach that does not require the knowledge of source node and sink prior to packet forwarding. The approach is developed for minimizing dark shaded areas and delays. Each node has the ability to calculate the flooding zone with the help of basic formula θ= π/2±α where α has the finite set of values between 0 and π/2. After computing flooding zone sensed data is flooded towards sink nodes. Sink node receives sensed data via upper layer nodes. Table VI shows the procedure of data packet forwarding.

The sender node N05show in Fig. 8 has a data packet ready to be sent with its L_ID. By using a simple Hello Packet (HP), the source node discovers candidate nodes within the flooding zone. Nodes that receive HP will reply to the source node with the help of Hello Reply (HR) which contains ID, L_ID, and its current Angle. The formats of HP and HR are illustrated in Table II and Table III. A node N14 resides in the flooding zone which will reply to source node N05 via HR. N05 will send the data packet to node N14 as its angle is closest to the vertical line as compared to other candidate nodes. The same procedure will be used by each node that wants to send data until it reaches the sink.

If multiple nodes reside within the flooding zone then each node will compute its angle and send it to the source node in the form of HR. the source node will compare the angle of each node, the data packet is sent to the node having the angle closest to the vertical line.

If the source node doesn't get any HR, the source node will increase the size of the flooding zone by incrementing the value of α until it meets the base condition 0 ≤ α < π/2. The nodes can use any random value of α because it is more useful for controlling power consumption and end to end delays. The values of α are dependent on the network. For sparse networks, the network uses higher values of α whereas, for dense networks, smaller values of α are used to adjust the size of the flooding zone.

In the worst case, if the source node has checked all values of α but does not get any HR from any node, then the source can forward the data packet to a node of the same depth level or at the horizontal line.

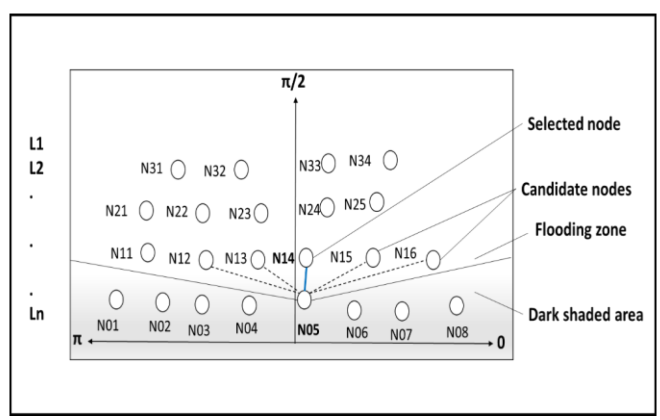


Fig. 8. Packet Transmission Process.

TABLE. VI. PSEUDOCODE FOR DATA PACKET FORWARDING

1. Initialize θ= π/2±α where 0≤α< π/2 // Here α is a variable
2. If Ns have DP //Ns is the sensor nodes and DP are the data packets
3. If 0≤α< π/2
4. Send HP to the Nn //HP are hello packets which are sent to neighbor nodes Nn
5. If HR received
6. Compare and store θ of Nc // Nc are candidate nodes for packet forwarding
7. If θ is close to Vl // check that if angle θ is close to Vertical line Vl
8. Consider it θa // consider that angle as adjusted angle θa
9. Send DP to Ns close to Vl
10. else go to step 6
11. else if α = π/2
12. Send DP to Hl // data packets will be sent to horizontal li
13. else increase Tr // if α is less than π/2 then transmission range is increased
14. Go to step 4
15. else go to step 17
16. Else go to step 17
17. End

In the rare case, if the source node can't find any node even on the same depth level, then the source node can eliminate the restrictions of the recommended communication range, which is 1km, and the node can directly send the data to the sink using maximum energy.

1) *Pseudocode for Packet Forwarding:* In the network when any node has a data packet ready to be sent which is either created by that node or is received from any other node, it firstly computes the flooding zone with the help of basic formula as discussed in section E. After computing the flooding zone, Hello Packets (HP) are flooded within it. The nodes which get hello packets will send Hello Reply (HR) in response to the hello packet along with their current angle. After receiving hello replies from all candidate nodes, the current angle of each node is compared. After comparison data is sent to the node having the angle closest to the Vertical Line (Vl). In case if the source node doesn't get any reply, the source node increases the flooding zone by incrementing the value of α. If no node is available till the maximum value of α, then the node can send data to Horizontal Line (Hl) or at the same depth level. The Pseducode and flowchart for Data Packet Forwarding is represented in Table VI and Fig. 9.

G. Implementation Tools

1) *Network Simulator (NS-3.26):* The development of simulation models that are sufficiently realistic for real-time network emulator, interconnected with the real world and allows many existing real-world protocol implementations to be reused.

2) *DESERT Framework:* It supports network simulator to design and implement of new underwater network protocols.

IV. RESULTS AND ANALYSIS

The chapter describes the simulation results to assess the performance of angle adjustment for diagonal and vertical communication which was presented in the previous chapter. Firstly, the simulation environment and criteria for assessment is described. The routing technique is assessed as per different parameters which involve node mobility and load on the network in Section C.

For the assessment of performance, AquaSim was used which is a network simulator based on NS-2 for the underwater environment. Aquasim supports node mobility, attenuation and 3-D deployment of nodes.

A. Simulation Environment

For the evaluation of the performance of the proposed approach NS-2 was used. Including the ordinary and sink nodes, a total of 300 nodes were deployed in a three-dimensional area of 800m×1000m2. Multiple sinks consisting of both acoustic and radio modems were deployed on the surface of the water.

The maximum distance among layers is 100m. The ordinary nodes which are deployed in the water were considered as floating between ground and surface of water whereas the sinks are static and placed on the surface of the water. Ordinary nodes can move horizontally with water in fixed motion up to 1-4m/s. The data can be delivered up to 9 layers from bottom to the surface.100m is the range of transmission of sensors. The average depth and width of layers is 100m. surface sinks are also deployed 100m away.

Consumption of power is different for different events. 1 unit of energy is required for transmission and 0.02 for receiving. For the prevention of packet collision, one node can transmit the packet at a time in the domain of collision. The MAC protocols are based on IEEE 802.11 with the Constant Bit Rate (CBR) of 512 bits per packet. 500 packets were produced at a rate of 1 packet/s. the anchored nodes produce 250 packets and the remaining were produced randomly by the floating node. Table VII presents the details of the simulation parameters. As a supplement tool, for plotting and analysis of graphs MATLAB was used.

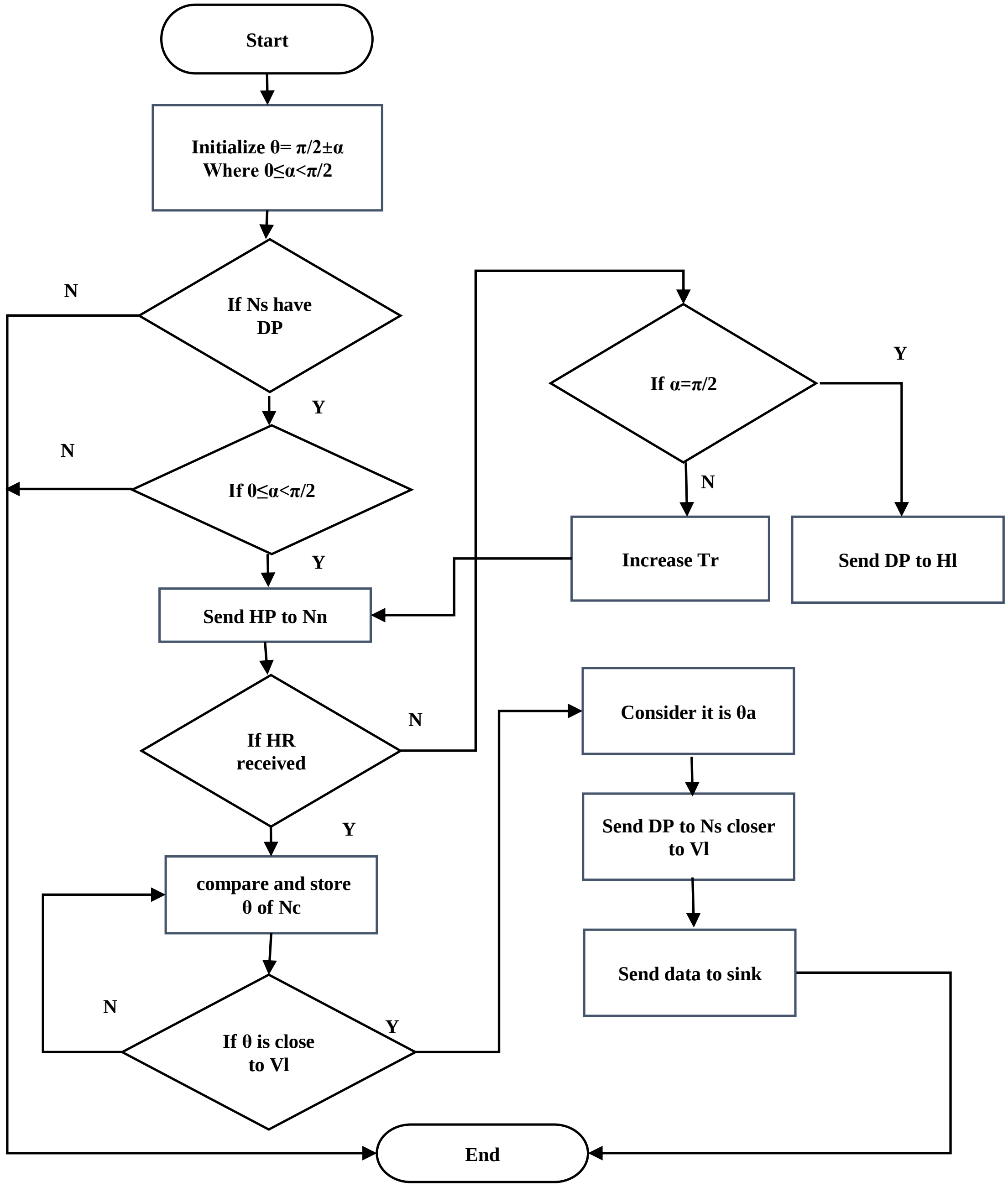


Fig. 9. Flow Chart for Data Packet Forwarding.

B. *Performance Metrics*

For the evaluation of performance proposed scheme consumption of energy, data delivery, and end-to-end delays were considered as evaluation parameters (as described in Table VII).

C. *Performance Analysis*

This section defines the evaluated performance of the presented approach based on the parameters which are node mobility, network load, and energy consumption.

1) *Mobility of nodes:* For data delivery ratio two-node movement speeds (2m/s, 4m/s) and static nodes were considered as illustrated in Fig. 10. With 300 nodes delivery ratio was 100% and it remains static. The delivery ratios don't face any major effects with the decrease of node density. If 30% of the nodes are not available still the network can achieve a 90% delivery ratio. Even if the network gets sparse and only 50% of nodes remain available, 85% delivery can be achieved. Only a minor difference was seen when the number of nodes decreases. But the difference was not high as 50% of nodes were available. Fig. 11 presents end-to-end delay and energy consumption with different node movements. A minor difference in the delay was noted at different speeds. This describes that no critical effect was on delays and energy consumption due to node movement.

TABLE. VII. SIMULATION PARAMETERS

Parameters	Values
Software Version	NS 2.33
Network Field	1500x800x800
Topology	3D
Simulation time/hr	6
Antenna	Directional (Parabolic)
Number of sensors	300
Number of sinks	8
Distance between layers	100 m
Transmission Range/m	100-150
MAC Protocol	IEEE 802.11
Packet Size	512 B
Bandwidth/Mbps	10
Packet rate/Kbps	6-16
Initial Node Energy/J	1000
Packet Transmission Energy/unit	1
Receiving Energy	0.02

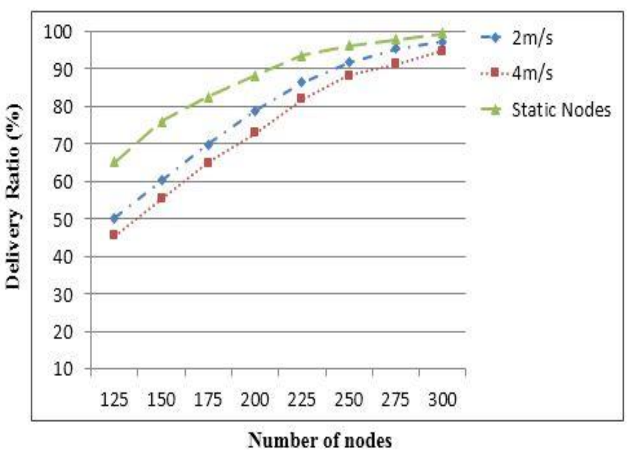


Fig. 10. Delivery Ratio for different Node Speeds.

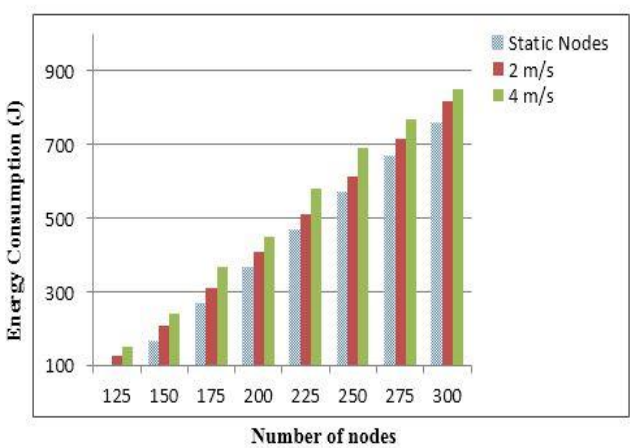


Fig. 11. Energy Consumption for different Node Speeds.

Mobility of node can affect average delays in sparse conditions, whereas all matrices show almost similar results in dense conditions.

Movement of the node doesn't affect the delivery ratio and energy consumption because complex routing tables were not maintained as per the node's location. Layer_ID of each node is maintained by node so its location can be easily handled.

Movement of the nodes don't affect the delivery ratio and energy consumption because complex routing tables were not maintained as per the node's location. Layer_ID of each node is maintained by node so its location can be easily handled.

2) *Network Performance with Different Number of Packets:* By generating one or more than one data packets, delays and delivery ratios of the proposed technique were analyzed so that the performance can be evaluated with different loads. Usually, 1 packet/sec is produced in the network but for non-normal cases, performance is assessed for 2 packets/sec and 3 packets/sec were also checked. The delivery ratio with different loads is shown in Fig. 12. The packet delivery ratios in the dense network were approximately the same. But, in sparse conditions, the difference appears where the load was high, but there were a smaller number of nodes in the network. Fig. 13 shows that the network can conveniently handles the situation when double packets are produced in the network, the end-to-end delays were still affordable.

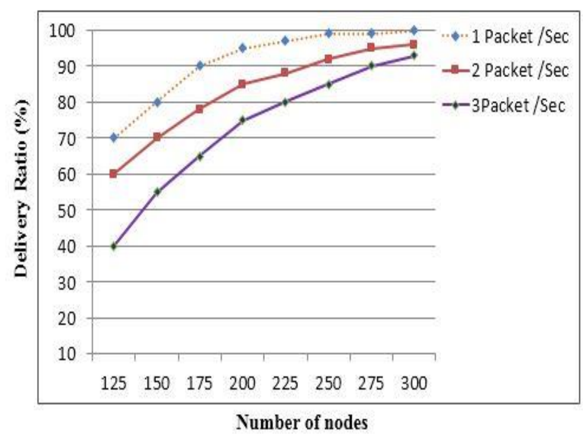


Fig. 12. Delivery Ratio for different Packet Loads.

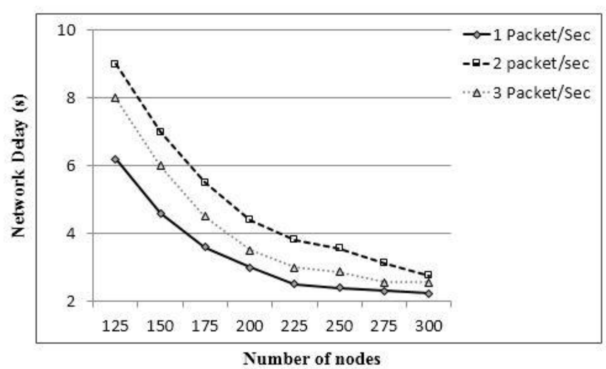


Fig. 13. End-to-End Delays for different Packet Loads.

3) *Performance Analysis with Different Flooding Zones:* Fig. 14 and Fig. 15 present the ratio of data delivery and end-to-end delays respectively with various values of α were selected randomly for analysis. Minor change can be observed with different values of variable α. The effect of using smaller and larger values of α for forwarding data packets can be seen in the dense network but the difference arises when the network becomes sparse and larger values of α are used to find a forwarder. However, the difference is obvious with less nodes. Still, the effect is affordable even with only 50% of available nodes.

D. Comparison with DBR

A lot of location-based routing protocols have been proposed. But in the water, it is quite difficult to get location information due to the unavailability of GPS [43][44]. Instead of using location information, DBR uses the depth of sensors for forwarding packets. Whenever a node has a data packet to send, it compares its current depth with the depth embedded in the packet. The packet is forwarded if the depth of the candidate node is less than the source node. Every node has the ability to calculate its current depth. The issue that DBR faces are: 1) more than one node can same data packet due to same depth which can cause power overhead. 2) DBR don't handle void area where no node is available which could be qualified as forwarder. Source node makes multiple tries but it doesn't select the route with higher depth due to greedy approach [45].

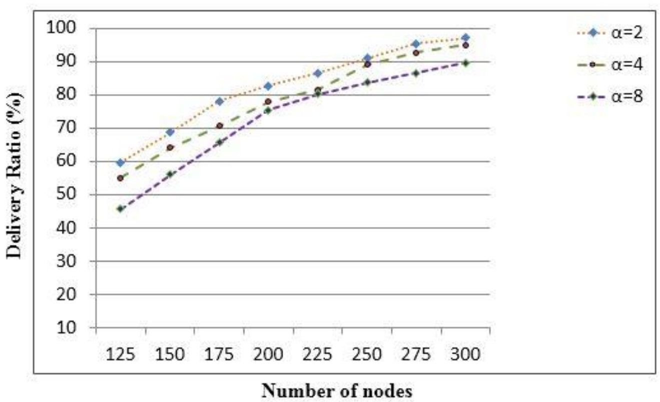


Fig. 14. Delivery Ratios for different Flooding Zones.

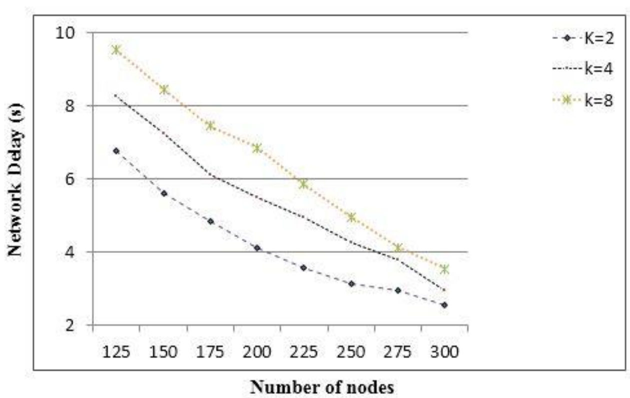


Fig. 15. End-to-End Delays for different Flooding Zones.

1) *Delivery ratio:* For evaluation of performance, delivery ratios of DBR and the proposed angle adjustment approach were compared in Fig. 16. The delivery ratio of both protocols decreases in sparse conditions but the effect on angle adjustment is less as it can send the data packet to the water surface without considering the location of sink nodes because multiple sinks are deployed unlike DBR [39].

2) *End-to-End delay:* The comparison of end-to-end delays among DBR and Angle adjustment are shown in Fig. 17. The holding time of DBR is responsible for delays whereas in angle adjustment approach node can directly flood the packets within the flooding zone.

3) *Energy consumption:* The power consumption of DBR and angle adjustment is presented in Fig. 18. The consumption of power is almost the same for a smaller number of nodes. But it increases in dense networks. The dense environment in DBR increases power consumption because it uses packet broadcasting and it also computes depth for selection of forwarder nodes, whereas angle adjustment only computes flooding zone for selecting next forwarder.

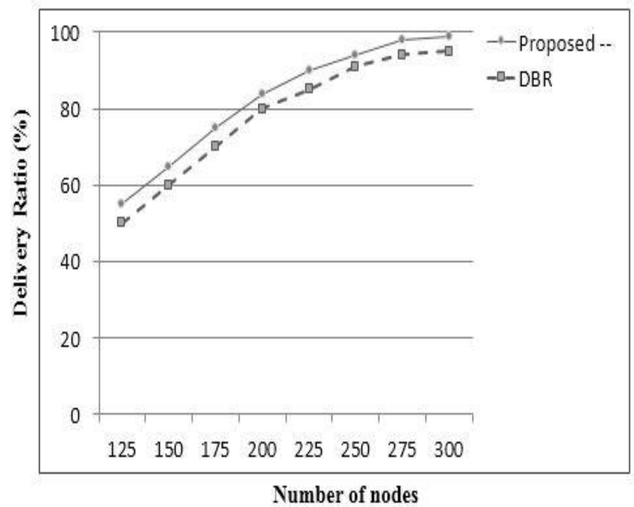


Fig. 16. Delivery Ratio Comparison with DBR.

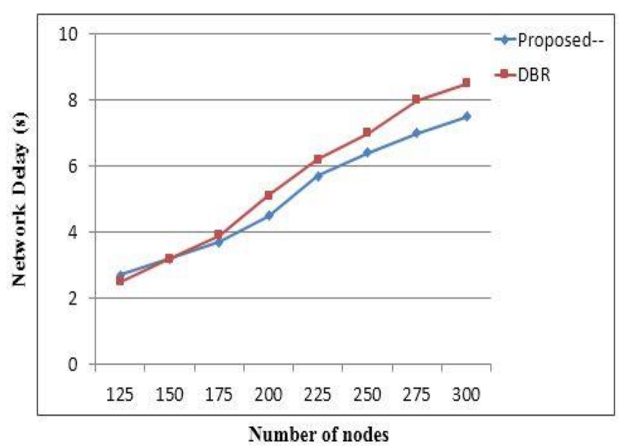


Fig. 17. Comparison of End-to-End Delays with DBR.

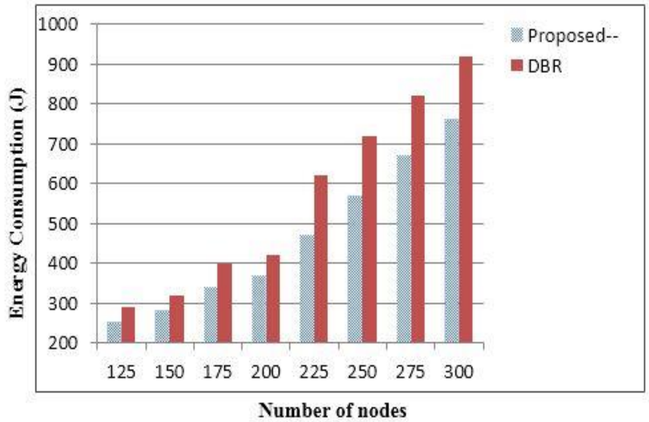


Fig. 18. Energy Consumption Comparison with DBR.

V. Conclusion

This research work aims to explore the problem of the dark shaded area within the environment of water. The solution for reducing the area is presented in the proposed approach. In this paper, an angle adjustment technique for diagonal and vertical communication is presented. It doesn't need the location information of nodes so don't maintain complex routing tables.

The approach is designed for selecting a fast routing path. It also reduces dark shaded areas, end-to-end delays, and horizontal communication. The approach considers all the values of angles between 0 and $\pi/2$ and changes the size of the flooding zone with each angle. The checking on each angle helps in finding the node which is closest to the vertical line. The vertical and diagonal communication reduces the long propagation delays.

Along with reducing the dark shaded area, the presented approach also achieved performance targets evaluated by using the NS-2 simulator with the AquaSim package. Simulation results reveal that the angle adjustment technique provides high ratios of data delivery, reduces end-to-end delays, and also decreases the consumption of energy.

The results of the angle adjustment technique were compared with DBR which shows that the proposed technique works well than DBR in terms of data delivery, end-to-end delays, and energy consumption in the dense as well as sparse network conditions.

References

[1] Zahoor, N. Javaid, M. Akbar, and Z. A. Khan, "A New Routing Protocol for Maximum Coverage in Square Field for Underwater WSNs," pp. 2–7, 2017.

[2] R. Diamant, R. Francescon, and M. Zorzi, "Topology-Efficient Discovery : A Topology Discovery Algorithm for Underwater Acoustic Networks," pp. 1–15, 2017.

[3] M. A. Kafi, J. B. E. N. Othman, and N. Badache, "A Survey on Reliability Protocols in Wireless Sensor Networks," vol. 50, no. 2, 2017.

[4] T. Ali, M. Ayaz, L. T. Jung, U. Draz, and A. Shaf, "Upward and Diagonal Data Packet Forwarding in Underwater Communication," vol. 1, no. 2, pp. 33–40, 2017.

[5] M. Khalid et al., "A survey of routing issues and associated protocols in underwater wireless sensor networks," J. Sensors, vol. 2017, 2017.

[6] J. Heidemann, W. Ye, J. Wills, A. Syed, and Y. Li, "Research Challenges and Applications for Underwater Sensor Networking," no. February 2015, 2006.

[7] B. Chen and D. Pompili, "Ad Hoc Networks Reliable geocasting for random-access underwater acoustic sensor networks," Ad Hoc Networks, vol. 21, pp. 134–146, 2014.

[8] A. Shaf, T. Ali, W. Farooq, U. Draz, and S. Yasin, "Comparison of DBR and L2-ABF Routing Protocols in Underwater Wireless Sensor Network."

[9] M. Khalid, Z. Ullah, N. Ahmad, and W. Khalid, "Comparison of Localization Free Routing Protocols in Underwater Wireless Sensor Networks," no. April, 2017.

[10] H. Yan, Z. J. Shi, and J. Cui, "DBR : Depth-Based Routing for Underwater," pp. 72–86, 2008.

[11] M. Ayaz, I. Baig, A. Abdullah, and I. Faye, "Journal of Network and Computer Applications A survey on routing techniques in underwater wireless sensor networks," J. Netw. Comput. Appl., vol. 34, no. 6, pp. 1908–1927, 2011.

[12] T. Ali, L. Tang, and I. Faye, "Diagonal and Vertical Routing Protocol for Underwater Wireless Sensor Network," Procedia - Soc. Behav. Sci., vol. 129, pp. 372–379, 2014.

[13] D. Kim, J. C. Cano, W. Wang, F. De Rango, and K. Hua, "Underwater Wireless Sensor Networks," Int. J. Distrib. Sens. Networks, vol. 10, no. 4, p. 860813, 2014.

[14] G. İşbitiren, "Three Dimensional Target Tracking With Underwater Acoustic Sensor Networks," vol. 60, no. 8, pp. 3897–3906, 2009.

[15] E. Felemban, F. K. Shaikh, U. M. Qureshi, A. A. Sheikh, and S. Bin Qaisar, "Underwater Sensor Network Applications : A Comprehensive Survey," vol. 2015, 2015.

[16] M. Ayaz and A. Abdullah, "Hop-by-Hop Dynamic Addressing Based (H 2 -DAB) Routing Protocol for Underwater Wireless Sensor Networks," 2009.

[17] N. Li, J. Martínez, J. Manuel, M. Chaus, and M. Eckert, "A Survey on Underwater Acoustic Sensor Network Routing Protocols," 2016.

[18] J. Preisig, "Acoustic propagation considerations for underwater acoustic communications network development," ACM SIGMOBILE Mob. Comput. Commun. Rev., vol. 11, no. 4, p. 2, 2008.

[19] K. M. Awan, P. A. Shah, K. Iqbal, S. Gillani, W. Ahmad, and Y. Nam, "Underwater Wireless Sensor Networks : A Review of Recent Issues and Challenges," vol. 2019, 2019.

[20] S. Tzeng and H. Lu, "QoS provisioning for multiple non-real-time services," no. February 2010, pp. 1392–1405, 2011.

[21] N. S. N. Ismail, L. A. Hussein, and S. H. S. Ariffin, "Analyzing the performance of acoustic channel in underwater wireless sensor network(UWSN)," AMS2010 Asia Model. Symp. 2010 - 4th Int. Conf. Math. Model. Comput. Simul., no. January 2010, pp. 550–555, 2010.

[22] I. F. Akyildiz, D. Pompili, and T. Melodia, "Challenges for efficient communication in underwater acoustic sensor networks," ACM SIGBED Rev., vol. 1, no. 2, pp. 3–8, 2007.

[23] I. F. Akyildiz, D. Pompili, and T. Melodia, "State of the Art in Protocol Research for Underwater Acoustic Sensor Networks *," 2006.

[24] I. F. Akyildiz, D. Pompili, and T. Melodia, "Underwater acoustic sensor networks: Research challenges," Ad Hoc Networks, vol. 3, no. 3, pp. 257–279, 2005.

[25] A. Manigopal, "UNDERWATER WIRELESS SENSOR NETWORKS : A SURVEY," vol. 2, no. 6, pp. 1254–1261, 2012.

[26] G. Xiang-ping, Y. yan, and H. Rong-lin, "Analyzing the Performance of Channel in Underwater Wireless Sensor Networks (UWSN)," Procedia Eng., vol. 15, pp. 95–99, 2011.

[27] N. Hemavathy, M. Mugesh, and P. Indumathi, "The Performance of Different Routing Protocols Available in Underwater Wireless Sensor Network-A Survey," no. 1, pp. 527–531, 2016.

[28] I. F. Akyildiz, D. Pompili, and T. Melodia, "Challenges for efficient communication in underwater acoustic sensor networks," ACM SIGBED Rev., vol. 1, no. 2, pp. 3–8, 2004.

[29] J. Premalatha and P. M. Joe Prathap, “A Survey on Underwater Wireless Sensor Networks: Progresses, Applications, and Challenges,” MATEC Web Conf., vol. 57, p. 02007, 2016.

[30] J. Kong, J. H. Cui, D. Wu, and M. Gerla, “Building underwater ad-hoc networks and sensor networks for large scale real-time aquatic applications,” Proc. - IEEE Mil. Commun. Conf. MILCOM, vol. 2005, no. November, 2005.

[31] M. K. Watfa, T. Nsouli, M. Al-Ayache, and O. Ayyash, “Reactive localization in underwater wireless sensor networks,” 2nd Int. Conf. Comput. Netw. Technol. ICCNT 2010, pp. 244–248, 2010.

[32] H. Nasir and N. Javaid, “Cooperative Routing in Underwater Wireless Sensor Networks (UWSNs) Cooperative Routing in Underwater Wireless Sensor Networks (UWSNs) By : Hina Nasir Supervised by : Dr . Nadeem Javaid Department of Computer Science,” no. July 2016, 2014.

[33] C. Uribe and W. Grote, “Radio communication model for underwater WSN,” 3rd Int. Conf. New Technol. Mobil. Secur. NTMS 2009, pp. 0–4, 2009.

[34] T. Kaur, “Underwater Wireless Sensor Networks Challenges : A Review,” pp. 1–5, 2018.

[35] J. Luo, L. Fan, S. Wu, and X. Yan, “Research on localization algorithms based on acoustic communication for underwater sensor networks,” Sensors (Switzerland), vol. 18, no. 1, pp. 1–25, 2018.

[36] M. Ahmed, M. Salleh, M. I. Channa, and M. F. Rohani, “Review on localization based routing protocols for underwater wireless sensor network,” Int. J. Electr. Comput. Eng., vol. 7, no. 1, pp. 536–541, 2017.

[37] X. Che, I. Wells, G. Dickers, P. Kear, and X. Gong, “Re-evaluation of RF electromagnetic communication in underwater sensor networks,” IEEE Commun. Mag., vol. 48, no. 12, pp. 143–151, 2010.

[38] T. Ali, L. T. Jung, and I. Faye, “Classification of Routing Algorithms in Volatile Environment of Underwater Wireless Sensor Networks,” vol. 6, no. 2, pp. 129–147, 2014.

[39] T. Ali, “Flooding Control by using Angle Based Cone for UWSNs,” pp. 112–117, 2012.

[40] T. Ali, L. Tang, and J. Ibrahima, “End-to-End Delay and Energy Efficient Routing Protocol for Underwater Wireless Sensor Networks,” 2014.

[41] Draz, U., Ali, T., Yasin, S., & Shaf, A. (2018, March). Evaluation based analysis of packet delivery ratio for AODV and DSR under UDP and TCP environment. In 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) (pp. 1-7). IEEE.

[42] Draz, U., Ali, T., & Yasin, S. (2018, November). Cloud Based Watchman Inlets for Flood Recovery System Using Wireless Sensor and Actor Networks. In 2018 IEEE 21st International Multi-Topic Conference (INMIC) (pp. 1-6). IEEE.

[43] Draz, U., Ali, T., Yasin, S., Naseer, N., & Waqas, U. (2018, February). A parametric performance evaluation of SMDBRP and AEDGRP routing protocols in underwater wireless sensor network for data transmission. In 2018 International Conference on Advancements in Computational Sciences (ICACS) (pp. 1-8). IEEE.

[44] Farooq, W., Ali, T., Shaf, A., UMAR, M., & Yasin, S. (2019). Atomic-shaped efficient delay and data gathering routing protocol for underwater wireless sensor networks. Turkish Journal of Electrical Engineering & Computer Sciences, 27(5), 3454-3469.

[45] Draz, U., Ali, T., Yasin, S., Fareed, A., & Shahbaz, M. (2019, July). Watchman-based Data Packet Forwarding Algorithm for Underwater Wireless Sensor and Actor Networks. In 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE) (pp. 1-7). IEEE.

[46] Yasin, S., Ali, T., Draz, U., & Rasheed, A. (2018). Simulation-based battery life prediction technique in wireless sensor networks. NFC IEFR Journal of Engineering and Scientific Research, 6, 166-172.

[47] Draz, M. U., Ali, T., Yasin, S., & Waqas, U. (2018, December). Towards formal modeling of hotspot issue by watch-man nodes in wireless sensor and actor network. In 2018 International conference on frontiers of information technology (FIT) (pp. 321-326). IEEE.

[48] Draz, U., Ali, T., Yasin, S., Waqas, U., & Rafiq, U. (2019, February). EADSA: Energy-Aware Distributed Sink Algorithm for Hotspot Problem in Wireless Sensor and Actor Networks. In 2019 International Conference on Engineering and Emerging Technologies (ICEET) (pp. 1-6). IEEE.

[49] Draz, U., Ali, T., Yasin, S., Waqas, U., & Rafiq, U. (2019, February). Towards Formalism of Link Failure Detection Algorithm for Wireless Sensor and Actor Networks. In 2019 International Conference on Engineering and Emerging Technologies (ICEET) (pp. 1-6). IEEE.

[50] Ali, T., Yasin, S., Draz, U., & Ayaz, M. (2019). Towards Formal Modeling of Subnet Based Hotspot Algorithm in Wireless Sensor Networks. Wireless Personal Communications, 107(4), 1573-1606.

Towards an Intelligent Approach for the Restitution of Physical Soil Parameters

Ibtissem HOSNI[1], Lilia BENNACEUR FARAH[2]
Laboratory of Remote Sensing and Information Systems with Spatial Reference, ENIT, Tunis El Manar University
Tunis, Tunisia

Imed Riadh FARAH[3]
Laboratory of Software Engineering Research
Distributed Applications, Decisional Systems and Intelligent Imaging ENSI, La Manouba University, Tunis, Tunisia

Raouf BENNACEUR[4]
Laboratory of Physics of Condensed Matter
Tunis, Tunisia

Mohamed Saber NACEUR[5]
Laboratory of Remote Sensing and Information Systems with Spatial Reference, ENIT, Tunis El Manar University
Tunis, Tunisia

***Abstract*—The analysis of the radar response on natural surfaces has been subject of intense research during the last decades in the field of remote sensing. Unless the availability of accurate values of surface roughness parameter, the restitution of soil moisture from radar backscattering signal can constantly provide inaccurate estimates. Characterization of soil roughness is not fully understood, so a wide range of roughness values can be obtained for the same studied surface when using different measurement methodologies. Various studies have shown a weak agreement between experimental measurements of soil physical parameters and theoretical values under natural conditions. Due to this nonlinearity and its ill-posedness, the inversion of backscattering radar signal on soils for restitution of physical soil parameters is particularly complex. The aim of the present work is the restitution of soil physical parameters from backscattered radar signal using an adapted backscattering model to the soil proposed description. As our study focuses on little rough soils, we have adopted in this work a multi-layered modified multiscale bi-dimensional Small Perturbation Model (2D MLS SPM). Subsequently, we propose a new way of describing the dielectric constant, with the aim of including air fractions in the multiscale multilayer description of the soil. Calculating the dielectric constant is based on the consideration of a soil comprising two phases, a fraction of soil, and an air fraction. For the inversion method, a methodology of coupling between neural networks (NN) and genetic algorithms (GA) was carried on in order to restitute the physical properties of the soil. Samples were generated by the original MLS 2D SPM followed by a neural network to obtain the statistic soil moisture and MLS roughness parameters algorithm. Thereafter, these restored values were modelled by the genetic algorithms to resolve, in part or in whole, the disagreement between the retrieval and original values.**

Keywords*—Inversion; air fractions; multi-layered; multiscale; SPM; genetic algorithms

I. Introduction and Background

Although soil moisture shows only a small proportion of the amount of water on Earth [12], it plays an extremely important role in different environmental sciences. This parameter, closely related to the soil dielectric constant, is strongly involved in the regulation of evapotranspiration phenomenon which directly affects atmospheric dynamics [10]. Modelling of this phenomenon is mainly related to a better understanding spatial organization structure of humidity.

Soil roughness plays an important role in the capture of water [13] [14] [15], promoting its infiltration and reducing its downward flow [1]. Therefore, the measurement of soil roughness should be taken into account to study and model the processes of runoff and erosion of agricultural land. It would also be very useful for better understanding the hydrodynamics and soils drainage.

Being two extremely dynamic variables under natural conditions, the possibility of achieving their estimations by means of remote sensing observations is very interesting for many applications [16].

Many studies have focused on the interest of radar remote sensing to characterize the soil and its parameters [11] [17] [18].

In this context, previous works [15], have characterized natural surfaces as stationary random processes where the distribution of heights defining the roughness of the soil is considered to be the superposition of a finite number of fractal one-dimensional Gaussian processes, each with a different spatial scale [8].

In the literature, a fractal surface model, where multiscale roughness is represented by two new parameters ν and $\gamma 0$ related respectively to the fractal dimension D and to the standard deviation of heights s [15], has been synthesized using the wavelet transform for one-dimensional surfaces.

However, to describe soil surfaces, one-dimensional profiles are insufficient because the roughness varies in all directions. Thus, in the context of this work, we have extended this multiscale description to bi-dimensional multilayered surfaces using the adapted bi-dimensional wavelet transform and the Mallat multi-resolution algorithm [2] [4], with the particularity of the presence of air pockets in the soil volume structure and its impact on the backscattering radar signal.

This paper is organized into two parts.

First part will carry the research methodology and it's organized into four sections.

The first section describes the bi-dimensional multiscale description of natural rough surfaces.

Section 2 shows the radar remote sensing opportunities for proposed natural soil characterization.

The third section deals with the direct problem, that is to study the SPM model adapted to the proposed description of the natural soil and calculate the received radar signal for a given atmosphere conditions and with a given surface geometry.

Section 4 discusses the opposite problem, which consists in estimating the geophysical soil parameters from remote sensing data obtained from the direct model.

The second part describes the results of the selected inversion algorithm in the restitution of physical soil parameters, namely the multiscale roughness and moisture, from the radar signal.

II. Research Methodology

A. Multiscale Description of a Natural Soil

A natural soil is generally described by its dielectric properties as well as its roughness which is directly related to the geometry of the surface. Conventionally, natural surfaces are described by two statistical parameters calculated from micro-topography profiles: a vertical component representing the elevation of the surface characteristics above and below its mean level, and a horizontal component explaining the lateral spacing between these characteristics. Generally, the first component corresponds to the amplitude and is expressed as a mean square deviation of heights (RMS for Root Mean Square), the second component represents horizontal variability and is associated with a correlation length (denoted L_C). The characterization of a surface only by one or the other parameter results in an incomplete description of the roughness. The larger RMS is, the greater vertical variations of the surface. The computation of L_C supposes to determine the autocorrelation function noted ACF, which expresses the correlation between pairs of measurement points separated by a distance Δ. It varies between -1 and 1 and, by definition, reaches its maximum for a distance $\Delta = 0$. It tends to decrease when Δ increases. The correlation length is then given by the value of Δ for which ACF fall to e^{-1} (about 37%) times its maximum value. The larger L_C is, the less the surface is showing horizontal variations.

Further characterization of roughness requires the consideration of its multiscale nature. Many studies [9] have shown that most natural surfaces can be described by a self-affine (or fractal) statistic according to fractal Brownian motion theory over a wide range of scales ranging from micrometer to kilometer [5]. In the case of a self-affine surface, an increase of the scale by an f factor in the horizontal direction results in a scale change f^H in the vertical direction (H is called the Hurst coefficient and quantifies the roughness change rate with the scale), and this to maintain the statistic characterization of the surface.

This assumption involves a single spatial scale for the surface characterization and its statistical properties by computing its standard deviation of heights and correlation length.

In this work, we consider the distribution of heights describing the soil roughness as the superposition of mono-dimensional Gaussian fractal processes each having a different spatial scale [6]. To describe this multiscale roughness in the case of bi-dimensional surfaces, we use the parameters ν and γ0 respectively related to the fractal dimension and the standard deviation of heights using the bi-dimensional wavelet transform and the Mallat multi-resolution algorithm [2] [4], to better describe natural surfaces.

B. Radar Backscattering Modeling by a Rough Surface

1) Proposed description of the studied soil: It's about describing the characteristics of a natural soil through its roughness and its dielectric permittivity. The rough surface, being considered stationary and ergodic in mean and variance, its roughness is entirely defined by the distribution and autocorrelation function of its heights. The second characteristic parameter is its water content. This is usually determined using the dielectric constant of the medium, a function of the soil moisture content, its composition, its temperature and the frequency of observation of the medium.

Conceptually, a multilayered approach was selected for the soil description and its reflection, since there are not really any physical layers but rather a continuous dielectric variability. A simplified technique was chosen for the volume. We proposed the use of an SPM surface diffusion model while considering the surface permittivity. We used for the upper layer a 2D multiscale description of soil roughness using the wavelet transform and the Mallat algorithm [2] [4]. The lower subsurface layer is divided into three fictitious layers separated by an assumed plane interface. We then calculated an effective resulting permittivity which includes the different dielectric permittivities of the three layers.

2) Introduction of air pockets in the soil volume structure: Because of the porosity characteristic of the soil, air can circulate through it, it's the aeration of the soil. More pores are numerous, well organized in network (good distribution between macro-pores and micro-pores, strong connections between the pores), more the air circulate easily towards the deep layers. A well oxygenated soil favors the mineralization of the organic matter in assimilable elements by the plant, and allows the breathing of the living organisms.

Interior works [7] modified the expression of the dielectric constant to take in consideration the presence of air pockets in the soil volume structure. The presence of air fractions is an influencing factor on relationship between soil moisture and the radar backscattering cross section from one study site to another and from one parcel to another.

3) Reflection of a multilayered medium: In this section, we propose the redefinition of the dielectric constant to include the air fractions presented in the soil volume structure, while taking into account the multiscale multilayer description of the soil.

The soil surface in considered as a three-layered medium, where D is the depth of radar signal penetration, as illustrated in Fig. 1. The multilayer soil model is composed of three uniform layers [8]: medium 0 is a half-space, medium 1 with a thickness $d1$ and a permittivity ε1, medium 2, which represents the soil air particles, has $d2$ as a thickness and ε2 as a permittivity, medium 3, which represents the soil layer below the depth of radar signal penetration D ($d1$ + $d2$), with a permittivity ε3: it's a semi-infinite layer that does not have a thickness.

Ei and Er are respectively the incident and reflected radar signal.

We have introduced the multilayer appearance of the soil surface moisture and thereafter, the dielectric constant will be redefined according to the new description to take into consideration the air / soil composition. ε_{app} is an effective permittivity that encompasses the different dielectric permittivities of the three layers.

$$\varepsilon_{app} = [v_{sol} * \varepsilon_{sol}^{\alpha} + (1 - v_{sol}) * \varepsilon_{air}^{\alpha}]^{\frac{1}{\alpha}} \tag{1}$$

α = 0.5, ε_{sol} is the dielectric constant of the soil, ε_{air} is the dielectric constant of air, and v_{sol} is the fraction of the soil defined as:

$$v_{sol} = -0.22\, Log(Z_s) + 0.0058 \tag{2}$$

$$Z_s = s^2/l \tag{3}$$

the volumetric water content is given by m_v :

$$m_v = -5.3\ 10^{-2} + 2.92\ 10^{-2}\varepsilon_{app} - 5.5\ 10^{-4}{\varepsilon_{app}}^2 + 4.3\ 10^{-6}{\varepsilon_{app}}^3 \tag{4}$$

C. *The Impact of Air Pockets Interfaces on RADAR Backscattering using Three Layered 2D MLS SPM*

To evaluate the effect of the air pockets in the soil, a sensitivity analysis of the backscattering signal is performed.

The backscattering coefficients expressions of three layered multiscale surfaces, for respectively the vertical and parallel polarization VV and HH are computed as follow:

$$\sigma_{vv} = 8k^4\sigma_1^2 \left| R_{\parallel} cos^2\theta + \frac{sin^2\theta(1+R_{\parallel})^2}{2}\left(1 - \frac{1}{\varepsilon_r}\right) \right|^2 W(2ksin\theta, 0) \tag{5}$$

$$\sigma_{hh} = 8k^4\sigma_1^2 |R_{\perp} cos^2\theta|^2 W(2ksin\theta, 0) \tag{6}$$

$$W^n(-2k_x, 0) = \frac{2}{\pi}\int_0^{\infty}\int_0^{\infty}\left(\frac{r_c^i(\xi,\eta)}{r_c^i(0,0)}\right)^n \cos(2k_x\xi)d\xi d\eta \tag{7}$$

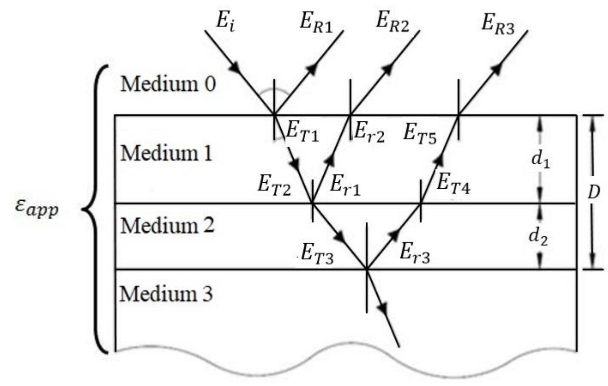


Fig. 1. Multilayered Soil Reflection Model [8].

θ is the incidence angle. W^n is the nth Fourier transform of the autocorrelation function given by Mattia, with n = 1 for the SPM model [15] [14] [3] [2].

To better show the adequacy of the three-layered 2D MLS SPM for the description of the soil, we present the table below (Table I) which shows a comparison between our proposed model and recent studies focusing on one and two layers:

The figures show the impact of the air pockets on the backscattering coefficient for the two polarizations VV and HH in an angular tendency from 20 to 70 degrees.

Fig. 2 and Fig. 3 show the impact of the ν parameter related to the fractal dimension on the radar signal backscattering, Fig. 4 and Fig. 5 present the impact of the γ parameter related to the standard deviation height on the radar signal backscattering, on both bare soil (solid line) and soil with air pockets (dotted line).

From these two figures (Fig. 2 and Fig. 3) presenting the impact of air pockets on the horizontal dimension of soil roughness ν, we can note that with an increase of the fractal dimension, the surface becomes smoother, and the value of the backscattering coefficient decreases due to the specular reflection.

Fig. 4 and Fig. 5 show the dependence of the vertical backscattering coefficient on soil roughness vertical dimension γ. The value of σ_{HH} increases with γ, since an increase in the standard deviation of heights can cause an amplification of the backscattered signal.

TABLE. I. COMPARISON BETWEEN MLS SPM WITH ONE LAYER, TWO LAYERS AND THREE LAYERS FOR ROUGHNESS AND MOISTURE PARAMETERS

	Single layered MLS SPM [19]	Two-layered MLS SPM [3]	Three-layered MLS SPM
Relative dielectric constant ε_r	ε_r	$\varepsilon_1, \varepsilon_2$	$\varepsilon_1, \varepsilon_2, \varepsilon_3 \geq \varepsilon_r$
Vertical roughness	γ_0 related to the standard deviation of the heights	γ_1, γ_2	γ_1, γ_2
Horizontal roughness	ν related to the fractal dimension D	ν_1, ν_2	ν_1, ν_2

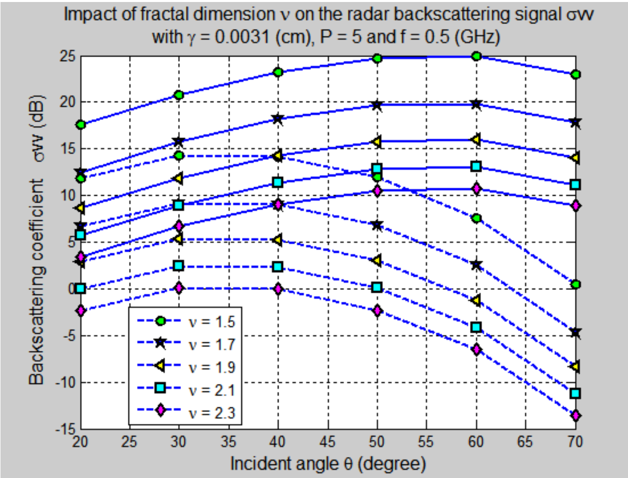


Fig. 2. σ_{VV} as a Function of Incident Angle θ (°) for different Values of ν with γ=0.0031 cm, P=5 and f=5 Ghz.

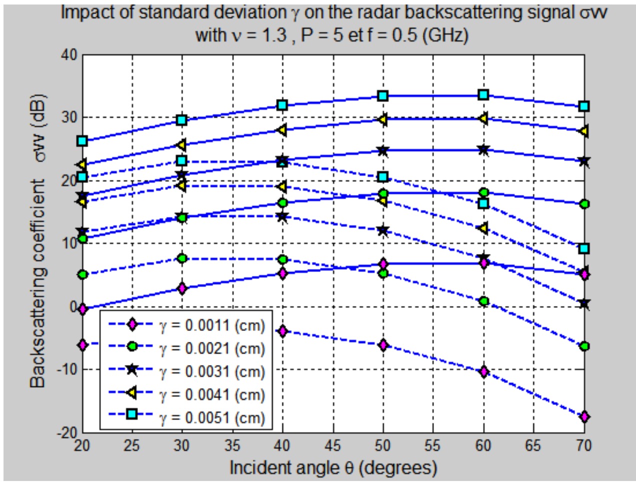


Fig. 4. σ_{VV} as a Function of Incident Angle θ (°) for different Values of γ with ν =1.3, P=5 and f=5 Ghz.

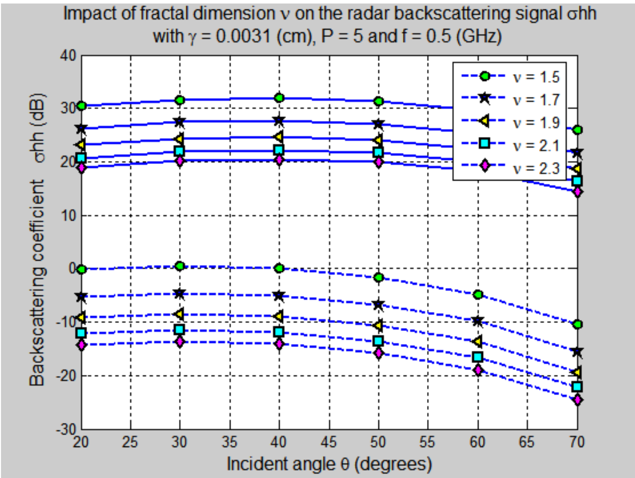


Fig. 3. σ_{HH} as a Function of Incident Angle θ (°) for different Values of ν with γ=0.0031 cm, P=5 and f=5 Ghz.

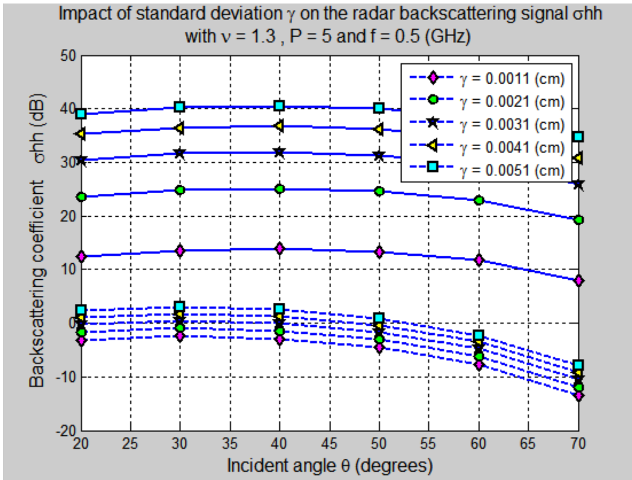


Fig. 5. σ_{HH} as a Function of Incident Angle θ (°) for different Values of γ with ν =1.3, P=5 and f=5 Ghz.

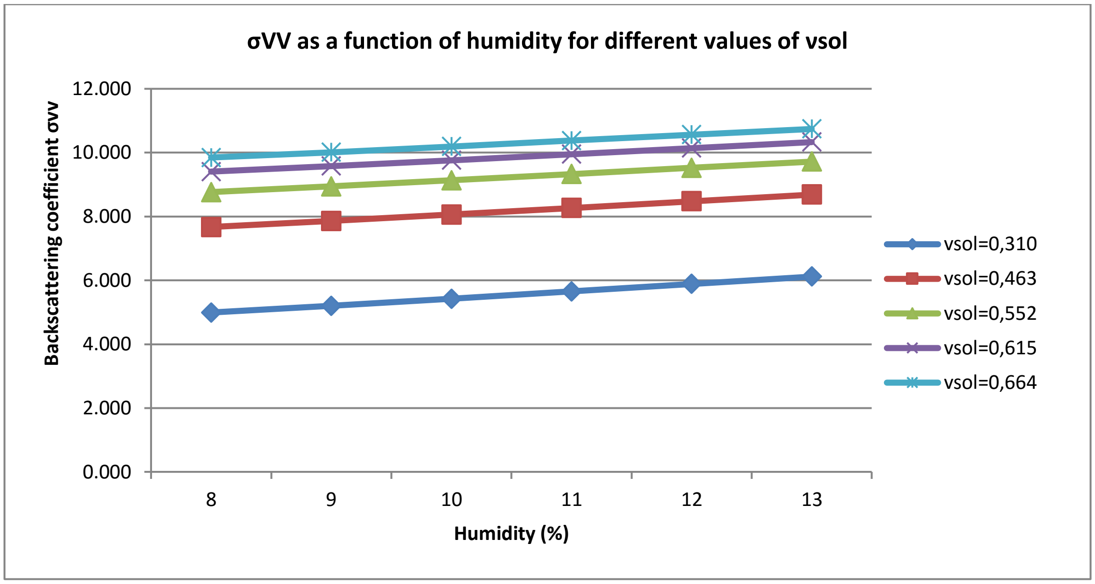


Fig. 6. σ_{VV} as a Function of Incident Angle θ (°) for different values of Humidity.

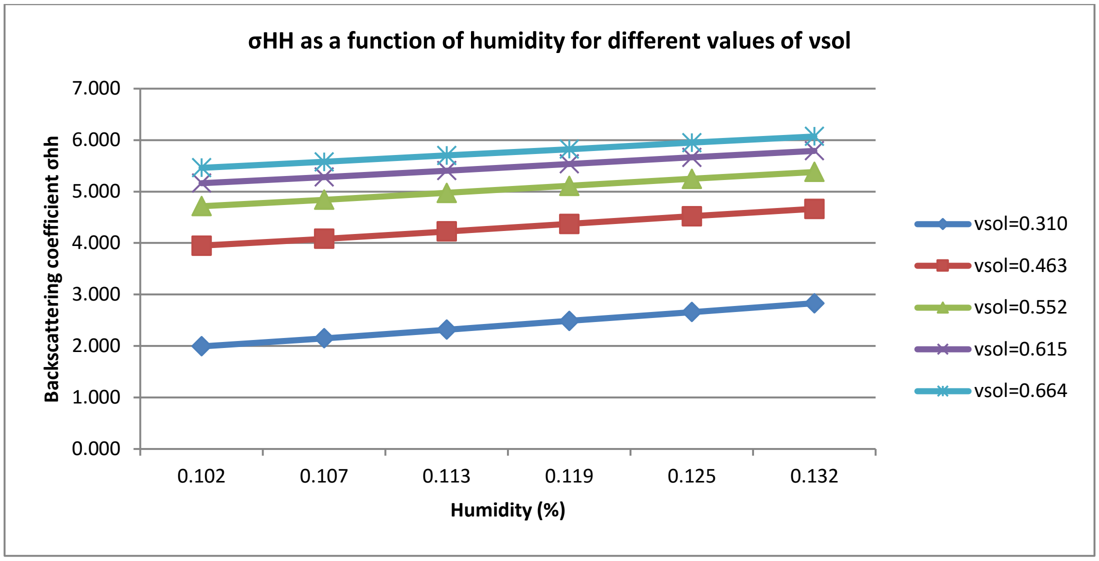


Fig. 7. σ_{HH} as a Function of Incident Angle θ (°) for different Values of Humidity.

In the following two figures (Fig. 6 and Fig. 7), we present the impact of air pockets in soil volume structure on the radar cross section.

Both figures (Fig. 6 and Fig. 7) show that with an increase in the soil moisture value, the amount of the backscattered signal will be higher and the penetration depth of the radar signal will be smaller.

We can note for all simulations, the amount of energy redirected to the radar system increases with the presence of air pockets in the soil texture. This increase reflects the inability areas containing air pockets to capture the emitted radiation and reflect it, and therefore a considerable portion of this radiation is returned to the radar.

D. Simulation of the Soil Physical Parameters with Volumic Air Pockets

Our objective is to find an inversion method to best restore the geophysical soil parameters. The number of retrieved soil parameters in the inversion procedure is limited to the number of the same parameters appearing in the direct radar signal simulation solution.

It is necessary to know how to deal with the direct problem, to have an algorithm calculating the bachscattered signal for a given atmosphere conditions and in a given surface geometry.

It is therefore clear that the calculation will be more reliable as the modeling of the backscattering radar signal will be more accurate. It is then necessary to use a precise and efficient inversion method that takes into account the information contained in the studied radar signal.

1) Neural network training: The main objective of the inversion algorithm consists in estimating the geophysical soil parameters namely the multiscale roughness $(\nu1, \nu2, \gamma1, \gamma2)$, and the multiscale moisture (ε_{app}) from the remote sensing data obtained from the direct model. This algorithm is based on neural networks method witch is trained by learning rules using the backpropagation method. Simulated data sets, based on the adapted SPM surface scattering model, are used to train the neural network (NN).

The inputs correspond to the observation data, while the outputs are the retrieved soil parameters. The adjustment of the internal parameters is done using a database that contains examples of inputs and outputs.

As shown in Fig. 8, backscattering coefficients were introduced as input parameters of the algorithm for different incidence angles measurements ranging from 20° to 70° for both horizontal and vertical polarizations (HH and VV).

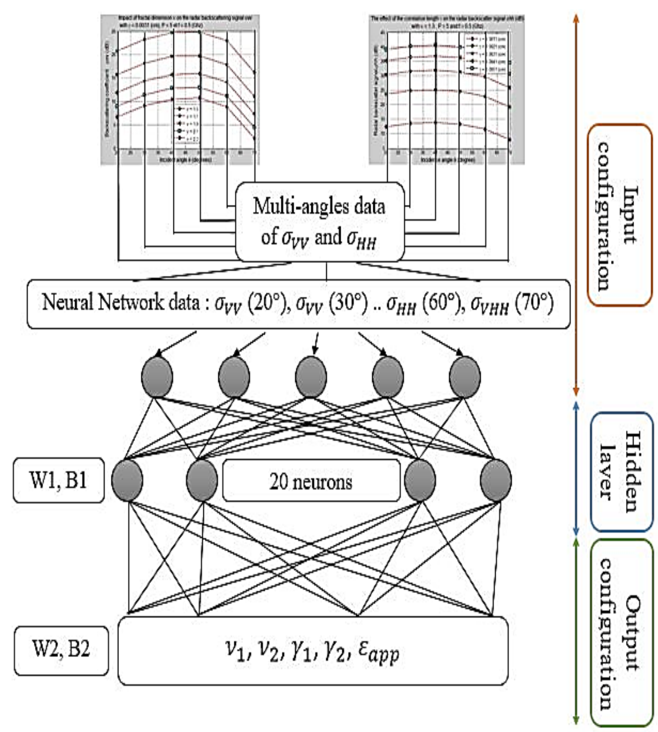


Fig. 8. Neural Network Inversion Method.

2) NN modelization by Gas: Generally evolutionary algorithms (EA), and genetic algorithms (GA) in particular, are directly derived from faculties of the nature to adapt to the environment by evolving through selection and reproduction. Neural networks (NN) are also a simplified way to simulate the abilities of organisms to adapt to their environment by learning. Simply because nature works well on this way, and successfully, it has been a source of inspiration for many works on hybridization of neural networks (NN) with evolutionary algorithms such as genetic algorithms, hoping that this combination can help resolve problems more effectively that the two methods taken independently.

The first step in allowing NNs to be handled by a GA is to define in which form in the data structure sense the EA will see the NN as an individual of a population (Fig. 9).

The evolution occurs on chromosomes that represent each of the individuals in a population. The process of natural selection ensures that the most suitable individuals reproduce more often and contribute more to future populations.

During reproduction, the information contained in the individuals of the parents is combined and mixed to produce the individuals of the children. Crossing result may in turn be changed by random perturbations.

For learning, we adopted an input layer with one hundred neurons (relative to populations with σ_{HH} , σ_{VV} and the incident angle θ), a hidden single-layer NN of twenty neurons, and an output layer of five neurons (relative to the parameters of soil moisture and roughness $\varepsilon_{app}, \gamma 1, \gamma 2, \nu 1, \nu 2$). These latter are derived from a set of weights that are trained with NN.

Both NN and NN-GA hybrid model have been implemented with the classical RPG (Retro-propagation gradient approach) to establish a comparison between them.

We start randomly by generating a population of individuals. To pass from a generation k to the generation k + 1, the following three operations are repeated for all the elements of the population k. Couples of parents P1 and P2 are selected according to their adaptations.

The crossing operator is applied to them with a probability Pc (generally around 0.6) and generates pairs of children C1 and C2. Other elements P are selected according to their adaptation.

The mutation operator is applied with probability Pm (Pm is generally much lower than Pc) and generates mutated individuals P0. The children (C1, C2) and the P0 mutated individuals are then evaluated before insertion into the new population. Different criteria for stopping the algorithm can be chosen:

- The wished number of generations to execute can be fixed from the beginning. This is what we are tempted to do when we have to find a solution in a limited time.

- The algorithm can be stopped when the population no longer evolves or is no longer fast enough.

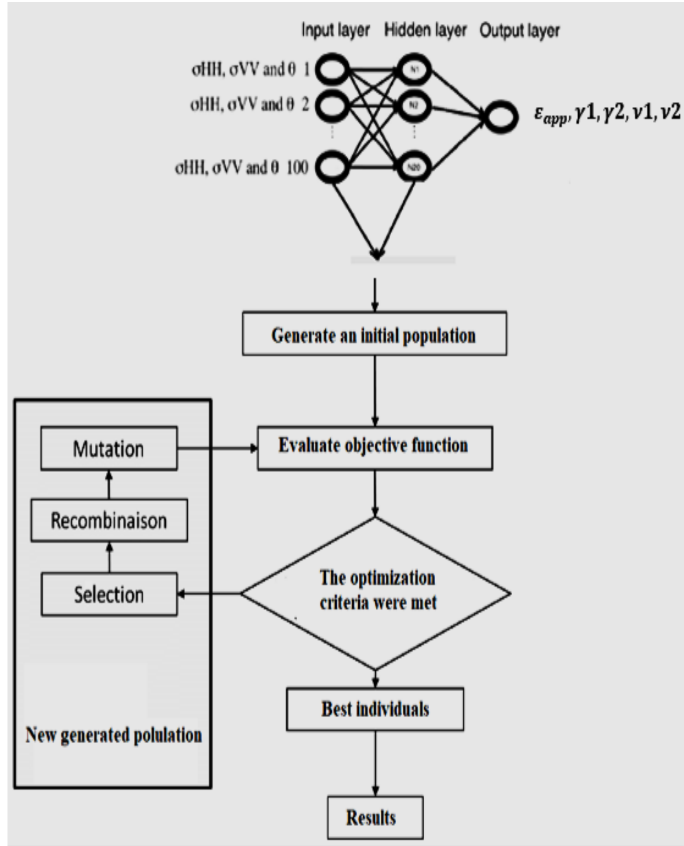


Fig. 9. Proposed Methodology.

The choice of the initial population strongly conditions the speed of the algorithm.

we start by Genome Coding, which is typically based on the storage of multilayer perceptron weights in the form of a matrix W such that w_{ij} corresponds to the weight of the connection from neuron j to the neuron i. Then simply put end to end each line of the matrix to obtain the genotype of an individual.

The biased roulette principle is used to randomly select pairs of reproducing individuals. This principle is based on the image of a roulette wheel such that the probability of selecting a particular individual is proportional to the value of its adaptation function. Thus, the best individuals will have a higher probability of being selected for reproduction.

3) Inversion algorithm results: The main objective of the inversion algorithm, which is based on a modeled NN by a GA method, is to retrieve physical soil parameters, namely the multiscale roughness and moisture, from the radar signal.

The obtained results with our approach NN-GA show that the retrieved parameters are improved compared to the inversion model based on NN. Fig. 10, Fig. 11 and Fig. 12 show that the retrieved dielectric constant with NN-GA are better than those obtained by NN.

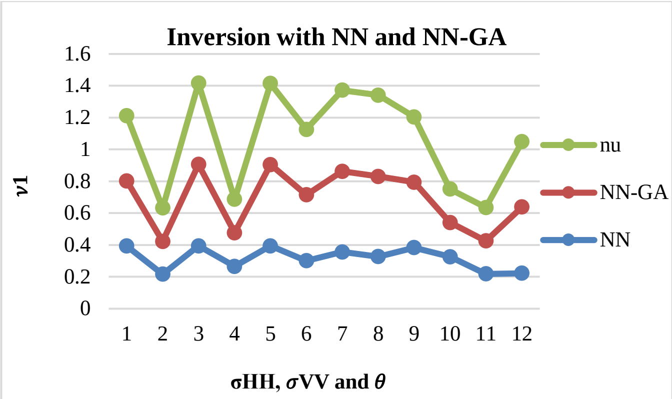


Fig. 10. Optimal Retrieved Fractal Parameter $\nu 1$.

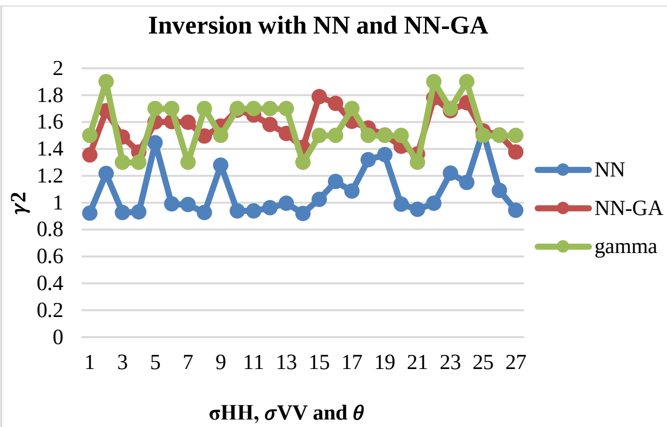


Fig. 11. Optimal Retrieved Standard Deviation γ2.

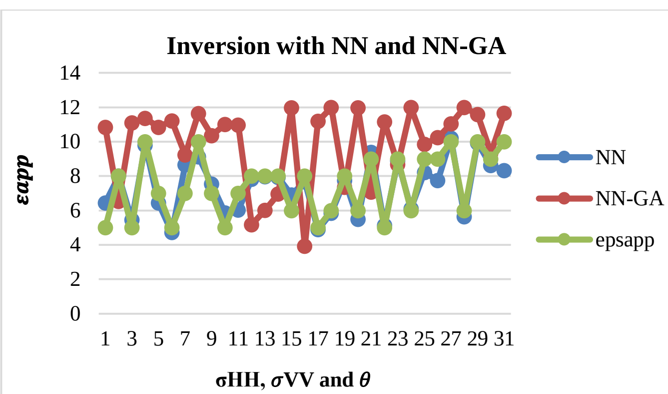


Fig. 12. Optimal Retrieved Dielectric Constant ε_{app}.

III. Conclusion

In this paper we have presented a synergistic method between the Neural Network and the modeled Neural Network by the genetic algorithms, to retrieve the physical parameters in a soil with volumetric air pockets.

The obtained results show that our proposed approach can, with a certain percentage of quadratic error, improve the retrieved parameters compared to the inversion model based on NN.

The main perspectives that follow the work carried out in this work are:

- Use other electromagnetic models in addition to the SPM model, the field of validity of which is restricted, for learning neural networks in order to be able to apply the neuronal inversion method to surfaces of any roughness.
- It will be interesting to apply the multiscale description on other types of surfaces such as maritime surfaces.

References

[1] N. Baghdadi, C. King, A. Bourguignon, A. Remond Potential of ERS and RADARSAT data for surface roughness monitoring over bare agricultural fields: application to catchments in Northern France Int. J. Remote Sens., 23 (2002), pp. 3427-3442.

[2] Farah, L. B, I. R. Farah, R. Bennaceur, Z. Belhadj, and M. R. Boussema, \Soil multiscale roughness parameters and soil moisture retrieval from radar backscattering using a neural network technique," PIERS Proceedings, 367-371, Tokyo, Japan, August 2-5, 2006.

[3] Farah, L. B., I. R. Farah, R. Bennaceur, I. Hosni, and M. R. Boussema, A two layers multiscale bi-dimensional SPM model for the study of radar backscatter behavior on semi-arid soil subsurfaces," PIERS Online, Vol. 6, No. 6, 509-513, 2010.

[4] Mallat, S.G. 1989. Theory of multi-resolution signal decomposition: The Wavelet representation. IEEE Transactions on Pattern analysis and machine intelligence. 1989. Vol. 11, 7, pp. 674-693.

[5] Vidal Vázquez, E., Vivas Miranda, J.G. & Paz González, A., (2005). Characterizing anisotropy and heterogeneity of soil surface microtopography using fractal models. Ecological Modelling, 182(3–4), pp.337–353.

[6] Ulaby, F. T., M. K. Moore, and A. K. Fung, Microwave Remote Sensing Active and Passive, 3, Artech House, Norwood, M.A., 1986.

[7] Zribi, M, Le Morvan, A et Baghdadi, N . 2008b. Dielectric Constant Modelling with Soil–Air Composition and Its Effect on SAR Radar Signal Backscattered over Soil Surface. Sensors. 2008b. 8, pp. 6810-6824.

[8] Song, K, Zhou, X et Fan, Y. 2009. Multilayer soil model for improvement of soil moisture estimation using the small perturbation method. Journal of Applied Remote Sensing. 2009. 3. 033567.

[9] Gorrab, A., Zribi, M., Baghdadi, N., Mougenot, B., and Chabaane, Z.L. (2015a). Potential of X-band TerraSAR-X and COSMO-SkyMed SAR data for the assessment of physical soil parameters. Remote Sens. 7, 747–766.

[10] Mohanty, B.P., Cosh, M.H., Lakshmi, V., and Montzka, C. (2017). Soil moisture remote sensing: State-of-the-science. Vadose Zone J. 16.

[11] Wang, H., Méric, S., Allain, S., and Pottier, E. (2014). Adaptation of Oh Model for soil parameters retrieval using multi-angular RADARSAT-2 datasets. J. Surv. Mapp. Eng. 2, 65–74.

[12] Chow V.T., Maidment D.R., Mays L.W. Applied hydrology. McGraw-Hill Inc.; New York: 1988.

[13] Wagner W., Blöschl G., Pampaloni P., Calvet J.-C., Bizzarri B., Wigneron J.-P., Kerr Y. Operational readiness of microwave remote sensing of soil moisture for hydrologic applications. Nordic Hydrology. 2007;38:1–20.

[14] Davidson M.W.J., Le Toan T., Mattia F., Satalino G., Manninen T., Borgeaud M. On the characterization of agricultural soil roughness for radar remote sensing studies. IEEE Trans. Geosci. Remote Sens. 2000;38:630–640.

[15] Mattia F., Le Toan T. Backscattering properties of multiscale rough surfaces. Journal of Electromagnetic Waves and Applications. 1999;13:491–526.

[16] Borrelli, P., D. A. Robinson Larissa, R. Fleischer, E. Lugato, C. Ballabio, C. Alewell, K. Meus-burger, S. Modugno, B. SchÄutt, V. Ferro, V. Bagarello, K. Van Oost, L. Montanarella, and P. Panagos, "An assessment of the global impact of 21st century land use change on soil

erosion", Nature Communications, Vol. 8, Article number: 2013, December 08, 2017.

[17] Oh, Y. 2004. Quantitative retrieval of soil moisture content and surface roughness from multipolarized radar observations of bare soil surfaces. IEEE Transactions on Geoscience and Remote Sensing. 2004. Vol. 42, 3, pp. 596 – 601.

[18] Srivastava, H.S, et al. 2003. Use of multiincidence angle of radarsat-1 SAR data to incorporate the effect of surface roughness in soil moisture estimation. IEEE Transactions on Geosciences and Remote Sensing. 2003. Vol. 41, 7, pp. 1638-1640.

[19] Bennaceur, L, Boussema, M.R et Belhadj, Z. 2005. Restitution des paramètres physiques de rugosité et d'humidité. 25 Juin 2005.

An Innovative Smartphone-based Solution for Traffic Rule Violation Detection

Waleed Alasmary
Computer Engineering Department
College of Computer and Information Systems
Umm Al-Qura University
Saudi Arabia

***Abstract*—This paper introduces a novel smartphone-based solution to detect different traffic rule violations using a variety of computer vision and networking technologies. We propose the use of smartphones as participatory sensors via their cameras to detect the moving and stationary objects (e.g., cars and lane markers) and understand the resulting driving and traffic violation of each object. We propose novel framework which uses a fast in-mobile traffic violation detector for rapid detection of traffic rule violation. After that, the smartphone transmits the data to the cloud where more powerful computer vision and machine learning operations are used to detect the traffic violation with a higher accuracy. We show that the proposed framework detection is very accurate by combining a) a Haar-like feature cascade detector at the in-mobile level, and b) a deep learning-based classifier, and support-vector machine-based classifiers in the cloud. The accuracy of the deep convolutional network is about 92% for true positive and 95% for true negative. The proposed framework demonstrates a potential for mobile-based traffic violation detection by especially by combining the information of accurate relative position and relative speed. Finally, we propose a real-time scheduling scheme in order to optimize the use of battery and real-time bandwidth of the users given partially known navigation information among the different users in the network, which us the real case. We show that the navigation information is very important in order to better utilize the battery and bandwidth for each user for a small number of users compared to the navigation trajectory length. That is, the utilization of the resources is directly related to the number of available participants, and the accuracy of navigation information.**

***Keywords*—Participatory sensing; traffic violation detection; automatic detection; applied computer vision; resources optimization**

I. Introduction

There are major challenges in transportation that require immediate innovative solutions [1]. The first one is road congestion. The second but most important one is accidents and fatalities. These are worldwide issues, where major cities and suburban areas become more congested, and car accidents with fatalities continue to occur. For example, in Saudi Arabia, there were 7000 fatalities and $45,000$ accidents in January 2015 [2]. It is clear that the disobeying traffic rules is among the top causes of fatalities, which motivated the government of the Kingdom to use smart stationary cameras equipped with sensors to detect the car speed on the highway, and also other cameras that could detect cars that crosses red light signal at intersections. However this solution is considered expensive. Using a system of *stationary* cameras requires manpower to operate. Therefore, we propose a smartphone-based system to detect traffic rule violation. To the best of our knowledge, this is the first paper that proposes such a system.

Current intelligent transportation solutions assume an on-board device within the car [3]. This device can monitor the location and speed of the car and then the *data* can be used to evaluate the traffic violation. A well-known product is provided by MobilEye that requires installing multiple devices and sensors into the car [4]. In [5], smartphones are used to detect the car turning or speeding, and it is intended to collect the data of the car that has the smartphone in it. In this paper, we propose a solution to detect the traffic rule violation of the other cars on the road, without the need to install any device in the *monitored* car, rather, we propose the use of smartphones as *monitoring* devices. We pursue this research with the mindset that technology can help in improving road congestion, and traffic accidents by combining driving traffic violation detection with participatory sensing.

Our proposed system automatically detects a number of traffic rule violations on the roads. Specifically, we propose that each driver willing to participate can download our application on his/her smartphone. The major sensor used is the camera in the device. This is different from the currently used technologies where an on-board unit monitor the vehicle in which it is installed. Our proposed system would make each car act as a monitor for the other vehicles on the roads.

The system has two main components, namely, an in-mobile component, and an in-cloud component. The in-mobile component first detects the traffic rule violation on the roads using based on multiple subsequent time frames. This component is fast and had limited resources, and hence limited accuracy. After that, if a traffic rule violation is detected, the snapshot of the video-frame is sent to the in-cloud component which is more intensive in resources and has higher accuracy.

We evaluate our proposed system by validation with video traces captured by cameras on the cars in the Middle East and USA. Our system shows a very high accuracy for multiple traffic scenario detection, which hypothetically represent traffic violation detection. We used multiple data sets for training and testing [6]–[21]. We used some images as is, and we created some data from video traces as will be explained in Section IV-A. Our results strongly supports the feasibility of the proposed scheme. Furthermore, we propose a model to optimize the use of participatory sensors selection under the limited resources scenario. The main challenge of this optimization problem is to simultaneously optimize multiple resources within a dynamic mobile environment. We show that the proposed model can achieve near-optimal resource utilization results compared to the fully known mobility dynamic.

The contributions of this paper are as follows:

- We propose a novel smartphone-based participatory sensing system for traffic rule violation detection that is accommodating to new computer vision, sensing, and networking technologies.
- We propose the relative positioning and relative speeding concept in out system design in order to infer the traffic violation from the computer vision algorithms and the smartphone sensors readings. These two metrics allow us to understand the driving context from which the traffic rule violation can be detected.
- We evaluate the proposed system using multiple data sets, and videos that are taken by smartphones (or smartphone like cameras).
- We propose a participatory sensor optimization framework to enhance battery and bandwidth utilization while guaranteeing smartphone sensing.

In the following section, we discuss the related works to our proposed framework. Then, in Section II, we explain the system model and the traffic detection framework, followed by explanation of the participatory sensor optimization scheme in Section III. After that Section IV demonstrates the experiment setup, the data set used for evaluation, and the evaluation results. We discuss the recent related works in the areas of applied computer vision and intelligent transportation systems in Section V. Finally, Section VI concludes the paper and outlines the position of the proposed framework within current and future technologies including autonomous vehicles.

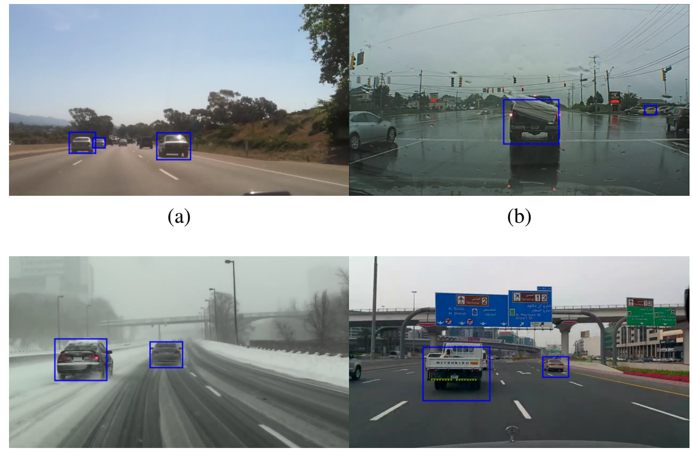
II. System Model

This section introduces an the overview of the proposed system, which is composed of three phases. The involved phases are as follows: 1) A fast in-mobile traffic rule violation detection phase, 2) A high-accuracy and cloud-based traffic rule violation detection phase, and 3) An in-mobile enhancement of the traffic rule violation detection phase. Each of the above mentioned phases is composed of a number of steps. Each of the following subsection describes the processes involved in each stage.

A. Fast in-mobile Traffic Violation Detection

This phase utilizes the captured video streams to detect the vehicles, the background, and the lane-markers. Then the said detected objects are used to track driving-behavior of all surrounding vehicles. This phase contains the following steps:

- Vehicles detection and classification.
- Lane marker detection.
- Motion detection.



(a) (b) (c) (d)

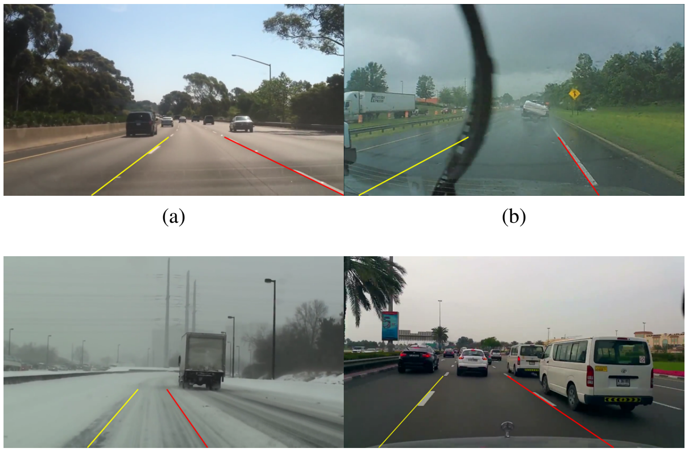
Fig. 1. Output samples of vehicles detection and classification using the in-mobile detector.

1) In-mobile: Vehicles Detection and Classification: We propose the use of the Haar-like feature-based cascade classifier driven by the AdaBoost algorithm as in [22] [23]. One reason is that this detector is very fast due to the use of integral images. It is also very accurate given a large number of training images, which is doable for such a crowd sourcing system. The main reason though is that optimized implementations are available for the algorithm (mainly designed for face detection), and can be integrated to the smartphone without much loss of speed. There are multiple variants and optimized detectors that also uses Haar-like features. We discuss them in Section V. A fast RCNN is the other candidate, but it is much slower than the Haar feature-based cascade classifier. It is shown in [24] that the processing time in CPU takes about 2 seconds, and the current smartphones are not powerful as our desktops. An RCNN can be used whenever the hardware allows it. This is the reason that deep learning is usually used in the server-side, for faster processing. Fig. 1 shows some samples of the outputs of this step. Each successfully detected vehicles is surrounded by a blue bonding box. [1]

2) Lane-markers detection: We use Hough transform to detect the two lane-markers on the street. We can also detect the two outer lane-markers on a three-lane, but we only focused on the two-lane-markers ahead of the car. This fast lane-detector was also used in [22], [23]. The lane-detector nominates lanes, and we only select the lanes if they are repeated in multiple frames. The repetition threshold can be adjusted as desired. We found that setting the repetition threshold to 4 or 5 is sufficient to detect lanes with high accuracy. ‘ Fig. 2 depicts some sample of lanes markers detection, the detected left and right lane markers are shown in yellow and red color, respectively.

3) Motion Detection: Motion detection can be performed using various methods. We chose our method to be aligned with the characteristics of the fast in-mobile detector. The main principle is as follows. Given an object that is detected around the same area in two consecutive frames in time, we map

[1]In Section IV-A, we explain which data set we used and the parameters used for training and testing of the classifier.



(a) (b) (c) (d)

Fig. 2. Output samples of lane markers detection of the in-mobile phase.

the features of the same object between the two frames. If all the mapped features move towards the top of the image, the object is assumed to be moving forward, and vice versa. Similarly, the object that is detected in the left side of the lane in one frame and mapped to an object that is detected in the right side of the same lane in a subsequent frame describes a motion from left to right, and vice versa. In this setting, we can detect multiple vehicle movements with respect to the available lanes. These motions can represent some traffic rule violations in some contexts. We use the SURF algorithm [25] for fast detection. We follow the feature matching with the variant of the random sample consensus (RANSAC), which is referred to as MSAC [26] and considered robust.

In order to make the motion detection procedure more accurate, we set a number of heuristic intuitive rules. First, we divide the image into three areas given the detected lane-markers. In this way, a detected vehicle bounding box must have its centroid somehow within in the correct position with respect to the lane-markers. Moreover, we also condition the far edges of the bottom points of the bounding box to be under the vanishing points of the image. These two procedures eliminate a good number of false positives. We can think of this detector as a motion inference tool. Combined with the information about the smartphone, it can understand the driving violation. For example, assume that the speed of the monitoring car is X. If the monitored car that is moving in front is speeding, the monitoring car can detect the speeding action speeding because the speed is $> X$. Another example is relating the camera parameters with my speed to detect the stalled vehicles. The detector also can infer extra information for approaching traffic light or a stop sign with high speed.

The false detection of the SURF algorithm are very minor. Moreover, when SURF is followed by MSAC, it provides matching points unless there is a very close match between the two different vehicles. Based on our experimental evaluation, we have not seen noticeable false detection. There are a few car detections on the background where there are objects similar to car structure. As the car is driving on the road, the background always moves backward. However, object motion is detected if all the matched points between the two consecutive frames are in the same direction, which is a strong condition for motion detection.

B. In-cloud Detector Traffic Violation Detection

The in-cloud detector is a component that is executed in the cloud. At the server side, we propose the use of a more powerful object detection and classification algorithms. We propose the use of the the deformable part-based models, and Discriminative learning with latent support vector machines (SVM) [27] [28] as an accurate detector. This detector is one of the most popular ones before the start of deep learning success in object detection. The detector has a very high accuracy and low recall rates, and can work on a regular CPU. We refer to this object detection method as Discriminatively Trained Part Based Models (DTDPM). This in-cloud detector has a processing delay that can take up to 1 second. *We performed multiple experiments on the laptop, and the processing delay is pretty consistent, but correlated with the image size*, but is very accurate in detecting vehicles, as we will demonstrate in the evaluation section. We also use the same lane-detection as in the fast detector. We also use convolutional neural networks (CNNs) to reduce the false positive detections. CNNs have several implementations and has a very high accuracy.

We also add an additional layer of verification using a *deep convolutional neural network* (CNN). Currently, there is a fast implementation of CNN [29] on the smartphone. However, it cannot perform object detection, rather, just classification. Therefore, our proposed in-mobile and in-cloud car detector extracts the possible vehicles and use a pre-trained CNN to remove false detection and affirm correct detections. We will demonstrate through the experiments that the accuracy of the CNN classifier is excellent. The CNN used in our model is based on [30], and has two classes, vehicles and backgrounds.

The lane-markers detection and motion detection is the same in both the fast in-mobile and in-cloud traffic rule violation detectors.

C. Data Collection and Dissemination Communication Models

For data communication, we define two data streams, data collection and data dissemination. The main focus in this paper is data collection. We assume that the data dissemination follows the current state-of-art technology. Data collection is mandatory, but can be performed in delay-sensitive or delay-tolerant methods. Sensors can opt in for delay-sensitive data collection, or they can use delay-tolerant option.

We assume two modes for data collection in our system model. First, we assume that the smartphone camera is used for data collection. In this mode, there is no communication overhead for data collection. Second, we assume that there are third party cameras that are used for data collection. These cameras can communicate with the smartphone via WiFi/Bluetooth. We assume that pairing is set up once, and repeated automatically every time the user enters the vehicle. In the sequel, we do not distinguish the two methods. We assume that the camera captured the data, and then the data is available at the smartphone for the next stage of processing.

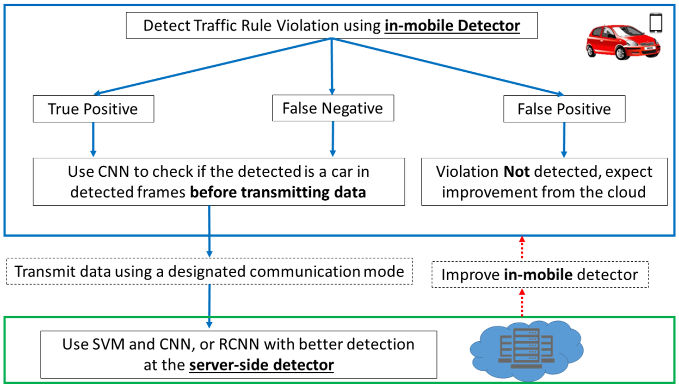
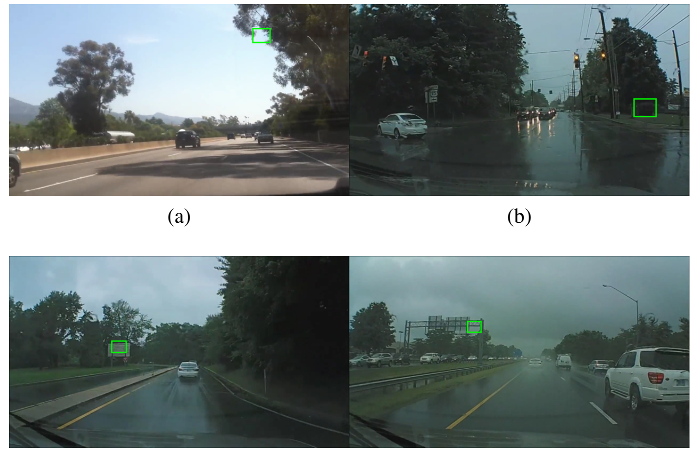


Fig. 3. Illustration of the Interconnection between the in-phone and in-cloud detection components.



(a) (b) (c) (d)

Fig. 4. Samples of the correct identification of the false positive detections using the CNN classifier. That is, incorrectly detecting backgrounds as vehicles. Those detections are then fed into the in-mobile detector.

Dissemination of collected data is according to the user preference and the available communication interface. We assume that our model can collect the data, and the dissemination is available according to the device communication capability. A user with DSRC communication device attached to his/her smartphone (e.g., mobile accessory by Arada systems DSRC-enabled devices [31]) can disseminate messages using vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), or vehicle-to-any (V2X) communication modes. A user who has a data subscription on the SIM card can transmit the data to the data centre, or disseminate warning messages using device-to-device (D2D) communication. The most convenient and low-cost option is to transmit data using WiFi offloading by treating the data in delay-tolerant communication mode.

D. Interconnection between the In-Mobile and In-cloud Components

The accuracy of different components of the system are previously known in certain applications. In this context, we know the limited computation capability of the smartphone compared to the cloud. Hence, we use the optimized detection at the cloud in order to improve the in-phone detector results. This can be performed in two ways. First, the negative results of the in-phone component can be neglected once the result is corrected at the cloud. Second, the neglected information can be used to improve the in-phone by retraining it. This also can be performed for improving the true positive detections as well. The interconnection between the two components is shown in Fig. 3. Fig. 4 shows sample outputs of correct classifications of false outputs detects by the CNN classifier using in-cloud detector. Those results are then fed for training in the in-mobile detector for better detection.

E. Driving Activity Detection

In this section, we aim at understanding the driving activity of each detected car on the road. The detected information is the respective position with every other detected object in the image. This information provides an understanding of each car on the road, the location with respect to the road lanes, and the relative motion with respect to the capturing phone. The general information we aim to collect in this section is illustrated in Fig. 5. The figure demonstrates how a car the road is divided into multiple lanes via the lane markers. The classifiers detect each car and then in subsequent frames detect their motion. We also show the regions of elimination of false positives.

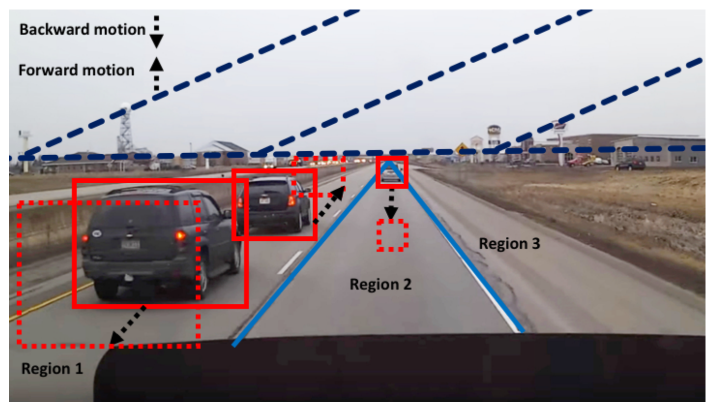


Fig. 5. The information detected by the driving activity detection model.

III. Participatory Sensing Optimization Problem

In this section, we assume that the smartphone parameters including battery and cellular infrastructure transmission are accessible by the our system. The user can also specify a quota that can be used by the application. The objective of the proposed optimization problem is to utilize the available resources to provide the best coverage of the roads to avoid missing any traffic rule violation while at the same time minimizing the battery usage and the data usage. We approach this problem in two stages in the following sections.

A. Sensor Selection Problem

Battery consumption is a critical parameter in participatory sensing. Many users might opt in to participate in improving the travel quality on the roads, reducing accident rates and fatalities, and capturing the traffic rule violators. However, the participants might refrain from using the proposed system because of the associated battery consumption with the application. Hence, we provide a custom solution based on each user preference.

It might be argued that an interested participant would use a charger in his/her car. That could be true. However, the fact that a monitoring app–even if optimized for battery consumption, would still use the battery that can be used for social networking or other entertaining apps is a critical point for discourage some users. Therefore, we optimize our model for battery consumption and participant efficiency. A participant is efficient in this context if his/her smartphone covers the intended areas of the transportation networks [2]. Therefore, a participant's efficiency in the sequel refers to the effective coverage of area.

It is critical to understand that scheduling of the smartphones based on their location is a well studied problem in the literature. The challenge in this context is that we only know the locations at the current position, and the car mobility can change in the future. That is, scheduling the participants have randomness in the future trajectories. Scheduling cars without taking into consideration the trajectory information of their future movement might result in lower spatial and temporal coverage, and under-utilization of the available battery and transmission bandwidth resources. Hence, we introduce the concept of an estimated trajectory-based scheduling. Therefore, the resources are not optimally utilized, but a near-optimal utilization can be achieved, which results in a practical solution.

We assume that based on the current location of the car, and its current speed and heading information, the future locations of the vehicles can be estimated within a short window of time. The estimation is not perfect, hence a slight degradation in the utilization of resources would be expected. Whenever the mobility information is not available anymore, the only solution that can be used is a non-optimal solution.

B. Battery Optimization Problem Formulation

Assume that some users do not want to use their app continuously. Therefore, we define B_i^Q as the battery quota that is allowed by user i for the crowdsourcing process. Suppose we have N participants. Let L be the number of participants where B_i^Q is unlimited. On the other hand, let M be the set of participants where B_i^Q is limited. M are the focus of our resource allocator. Note that if N can provide the required spatial and temporal coverage of roads, then we do not have to use the M participants. However, we do not consider this case in our study. We consider the case where we are forced use the M resource limited participants. We assume that the location of all participants in the system are known at the scheduling time. Let $a_i[t]$ denote the activity of fast detector on the smartphone i at time t. That is

$$a_i[t] = \begin{cases} 1 & \text{if fast detector is activated} \\ 0 & \text{otherwise} \end{cases}.$$

[2]Accuracy is mainly related to the computer vision algorithms discussed in the previous sections, and is not considered as a metric here.

Then selection among the M participants to run the fast detector is performed via the following optimization problem in a centralized fashion[3]:

$$\begin{aligned} \underset{a_i[t]}{\text{Minimize}} \quad & \sum_{i=1}^{M}\sum_{t=1}^{T} a_i[t] && (1)\\ \text{subject to} \quad & \sum_{i=1}^{M} a_i[t] \odot b_i[t] = 1,\ \forall t \in \{1,\cdots,T\}, && (2)\\ & \sum_{t=1}^{T} a_i[t] \le B_i^Q,\ \ \forall i \in \{1,\cdots,M\}, && (3)\\ & a_i[t] \in \{0,1\}, b_i[t] \in \{0,1\}. && (4) \end{aligned}$$

where $b_i[t]$ is an indicator that is set to 1 if the mobility is estimated in the near future; otherwise is set to 0[4]. Moreover, $B_i^Q < B_i$ is the amount of battery that is allocated for the fast detector at one scheduling epochs. In other words, the user specify $B_i = kB_i^Q$, where k is a predetermined parameter that demonstrates the length of acceptable predictability of mobility information. The above optimization problem allocates only one participant to cover an uncovered area for each time instant (Constraint 2). Moreover, it schedules sensors such as the battery consumed over time is related to the allocated battery consumption by the participant (constraint 3). Furthermore, it allocates participants according to their predicted availability in the uncovered locations using the variable $b_i[t]$ (constraint 2).

C. Realtime Transmission Problem Formulation

For the offloading communication option, all participating participants can collect data and transmit them using WiFi, whenever there is an available connection. However, when delay-sensitive data transmission is limited by the user, we optimize the transmission according to the provided quota set by each user. We also provide a single transmission for each captured event among all users. Let

$$T_{x_i[t]} = \begin{cases} 1 & \text{for delay-sensitive transmission} \\ 0 & \text{otherwise} \end{cases},$$

and D_i^Q, the normalized number of transmission times of captured traffic violations using the fast detector. Normally, the fast detector is used, then the transmission is established. Therefore, the delay-sensitive transmission scheduling is performed by the following scheduler:

$$\begin{aligned} \underset{a_i[t]}{\text{Minimize}} \quad & \sum_{i=1}^{L}\sum_{t=1}^{T} T_{x_i[t]} && (5)\\ \text{subject to} \quad & \sum_{i=1}^{L} T_{x_i[t]} b_i[t] = 1,\ \forall t \in \{1,\cdots,T\}, && (6)\\ & \sum_{t=1}^{T} T_{x_i[t]} \le D_i^Q,\ \forall i \in \{1,\cdots,M\}, && (7)\\ & T_{x_i[t]} \in \{0,1\}, b_i[t] \in \{0,1\}. && (8) \end{aligned}$$

[3]The central controller can be the fusion center or a cluster-head node within a group of nodes.

[4]Whenever $b_i[t] = 0$, the scheduler loses location information and would results in under-utilization of the resources.

This optimization problem would results in transmitting only one detected traffic violation in the same area (Constraint 6). The transmission quota of each participant is satisfied in Constraint 7. Note that we do not perform any battery constraint in the transmission. However, adding a battery constraint for transmission is doable. It should add a parameter that will constraint the number of transmissions. We assume that D_i^Q includes that information. In other words, the user enters the number of transmission time quota, and we calculate D_i^Q accordingly.

D. Battery and Transmission Optimization

It might be argued that problems (1)-(4) and (5)-(8) can be performed as a single optimization problem. However, the concept in our crowd sourcing system is that each user might chose to fully utilize activation, but not transmission, or vice versa. In other words, capturing an event does not necessarily require a delay-sensitive transmission. In fact, it might be expected that most users of such a crowd sourcing system will perform sensing without any delay-sensitive transmission as long as the transmission over the cellular network is not free. The incentive model that we use in the sequel does not require the users to pay extra money, unless they are willing to. Therefore, covering the traffic violation of the roads is the main objective of this work whether a realtime transmission occurs or not (realtime transmission of vehicular information as a separate topic that is widely researched). In addition to that, transmission might affect the battery consumption especially over cellular infrastructure. This directly reflects in our problem where we actually excluded the group of workers who are offering limited battery consumption. Adding that group of workers would result in chaining the parameters in problem (5)-(7), but not the scheduler design.

Therefore, we also add a third optimization problem, where each activated camera is activating the fast detector and transmitting given the fact it has a battery limitation and transmission budget. Let $c_i[t] = a_i[t] \odot T_{xi}[t]$, where $\odot$ is the binary product. Then the combined battery and transmission optimization problem becomes

$$\underset{c_i[t]}{\text{Minimize}} \quad \sum_{i=1}^{M} \sum_{t=1}^{T} c_i[t] \tag{9}$$

$$\text{subject to} \quad \sum_{i=1}^{M} c_i[t] \odot b_i[t] = 1, \ \forall t \in \{1, \cdots, T\}, \tag{10}$$

$$\sum_{t=1}^{T} c_i[t] \le B_i^Q, \ \ \forall i \in \{1, \cdots, M\}, \tag{11}$$

$$\sum_{t=1}^{T} c_i[t] \le D_i^Q, \ \ \forall i \in \{1, \cdots, M\}, \tag{12}$$

$$c_i[t] \in \{0, 1\}, b_i[t] \in \{0, 1\}. \tag{13}$$

This optimization problem actually combines the two set of constraints (2)-(4) and (6)-(8), which results in combining the fast detector activation and transmission according to the availability of battery power and transmission quota.

TABLE I. List of data sets used for training the classifiers and vehicle detectors.

System Component	Training Data	Size
Fast detector (Positive)	Caltech car data set [6], [7]	1,182
Fast detector (Negative)	Caltech background data set [8]	1,599
CNN detector (Positive)	Extracted cars from [14], [15]	1,186
CNN detector (Negative)	Caltech background data set [8]	1,599

IV. Performance Evaluation

A. Experimental Setup

The proposed framework can be tested on a smartphone on the roads. However, such an experiment would consume extra time and effort for collecting the data and testing. Therefore, we chose to setup our evaluation procedure as follows. First, we downloaded several vehicles and backgrounds data sets from internet for training and some testing [6]–[13]. Second, we downloaded different videos from YouTube that are taken by smartphones and are used to capture the road [14]–[21]. The experiment we performed uses two separate training and testing sets for each part of the system. In other words, we used different training sets for different parts of the system, and the testing data is different from those data sets. A detailed information on how we trained the data is shown in Table I. For the positive training of the fast detector, we use the extended data that can be found at [7]. For the negative samples of the fast detector, we used Caltech background data set and added some negative images from [9]. For the CNN detector, we used extracted positive training samples from two videos using the car detector in [27], [28], and then removed images that have large background potions. For the negative samples, we removed any background images in [8] that contains cars in them. For the CNN detector, we used transfer learning of the pre-trained network [30].

For testing the CNN detector, we used the positive samples that contains cars in data set in [10]. We removed the background, and just used the bounding box of the car. We used the negative background samples in [11] for background detection. Note that the testing data were not used in the training of the fast detector nor the CNN detector at all. The CNN network has two outputs indicting whether the image is a car or not. For the testing of the system framework, we use the videos in [16]–[21]. Neither the videos nor images extracted from the videos were used for training at all. The videos represent different cities, different driving conditions including sunny, during and after sunset, night time, raining time, and snowing time.

The design of out experiment is based on the currently available data sets and videos. We believe that the accuracy of our proposed model can be enhanced given larger number of sets, which can be though a large-scale crowd sourcing or deployment of the system.

B. Performance Metrics

To evaluate the proposed system, define two categories of metrics, namely, accuracy metrics, and efficiency metrics. The accuracy metrics are meant to evaluate the accuracy of the components of the driver detector. The efficiency are meant to evaluate the efficiency of the components of the framework including the driving detectors, and the resource scheduler.

Accuracy Metrics:

- *Accuracy*: The accuracy of the traffic violation detection is directly related to the accuracy of the computer vision algorithms. We define the accuracy as the percentage of correct classification using the machine learning algorithm.
- *Number of detected objects*: We use the number of identified objects as an indicator for the machine learning algorithm. It provides an intuition on the accuracy of the algorithm when we compare different algorithms with the same data set.
- *Above Horizon*: The horizon in our experiment is the vanishing point intersecting the two detected lanes. It is a measure of false positive vehicle detection.

Efficiency Metrics:

- *Processing time*: The time it takes to complete a specific computer vision operation.
- *Battery usage metric*: We define battery usage metric as the number of times the battery is used within the assigned budget by the participant, divided by the scheduling time. The lower this metric reflects a better use of the battery due to a better scheduling of the workers. Having a smaller battery usage metric during a scheduling epoch prolongs the use time of the each participant, and hence tends to provide more coverage on the roads.
- *Bandwidth usage metric*: Similar to the battery usage metric, the bandwidth usage metric is the number of times a transmission occurs given the assigned transmission budget, divided by the scheduling time. A lower bandwidth usage metric indicate better scheduling of the workers in favour of transmission budget.

Traffic Violation Metrics:

- *Detected motion*: This is the motion of the monitored vehicle. It can be detected as moving forward or moving backward between subsequent frames.
- *Relative position with respect to lanes*: This is the relative position of the detected cars with respect to each of the detected lanes. In other words, we know if the car is to the right of a lane or to the left.

C. Experimental Results

1) CNN Classifier: The CNN classifier reached very small incorrect classifications based on our training and testing. In order to show that, we test on labeled data of cars and backgrounds, and we show the results in Table II. We have tested the data over $16,186$ car images from [10], and the accuracy is more than 92%. We have also testing the data over $1,156$ backgrounds from [11], and the accuracy is more than 95%. We believe that the training data could be improved in order to generalize the system to get a better accuracy. We also notice in the same figure that the processing time is large due to the fact of using CNN on Matlab, and Matlab is known to be slow, and the implementation of the CNN is currently not optimized for speed. However, for realtime implantation, SDKs such as [29] can be used. A recently optimized implementation for speed has been published achieves an average precision of at most 88.7% on smartphones [32].

TABLE II. THE ACCURACY AND PROCESSING TIME FOR THE CNN CLASSIFIER.

Classification	Accuracy	Processing Time
Car	92%	1.3799 seconds (on CPU)
Background	95%	1.3763 seconds (on CPU)

2) Single Video Resolution for Training and Testing: We now demonstrate the accuracy metrics of the proposed framework on average. We use the video from YouTube [16]. We run the algorithm over the videos and plot the performance metrics in Fig. 7(a), (b), and (c). We use one video to show the relative performance between the fast and in-cloud detectors. In Fig. 7(a), we the number of frames fed to each detector, the number of detected cars, and the results of post processing after detection. We can see the number of detected cars is somehow close between the two detectors. However, after pre-processing, more than $2/3rd$ of the vehicles that were detected by the fast detector are eliminated due to false detections, detection of elements above the horizon, or the bounding box is not properly aligned between the detected lane markers. We noticed that if we relax the lane markers constraint, then the number of vehicles after post-processing becomes closer to the in-cloud detector. It is also expected that the in-cloud detector is more accurate especially in eliminating the false negatives. However, to minimize the number of false detections in the fast detector, we used the CNN detector and post-processing.

In Fig. 7(b), we show the number of detected motions. That is, out of the multiple detected cars, in which a small number of the detections might be false, the number of relative motions is much smaller than the number of vehicles. This is due to the fact that a detected motion has to be between two images of the same vehicles, and the feature matching-based motion detection is very strict in favour of accuracy. The accuracy arises from the fact that we only detect a vehicle motion if all matched feature points move in the same direction. This procedure eliminates the noise in the background (i.e., another vehicle detected within the bounding box, or the sidewalks). We can see that the number of matching in in-cloud detector is lower than the fast detector. We checked multiple frames in the in-cloud detector, and we found that the detector usually provides a bounding box that contains a noticeable portion of the background. Features matched to the background will reduce the accuracy of the motion detector. The same figure shows the number of detected lane marker is the same for both detectors, which is normal for a Hough transform-based lane-marker detector. However, the number of detected relative positions with respect to the lane marker is much lower than the number of the detected cars using the either fast or in-cloud detectors. This is due to the fact that we do not consider a lane-marker until it gets repeated for multiple consecutive frames, and that we require the centroid and the edges of the bounding box to be in the same relative position to the lane-marker. Finally, the figure shows that we detected 543 false positive detections above the horizon. This is how post-processing eliminates such false detections. The in-cloud detector did not results in any cars that are positioned above the horizon.

We show the processing times for each detector including the lane marker detector in Fig. 7(c). We can see that the fast detector is very fast while the in-cloud detector takes about 2 seconds (0.511×4 as the results are scaled in the figure for the first column). The lane marker detection is very fast, too.

3) Different Video Resolution for Training and Testing: What we want to achieve in this section is detection of the traffic violation with multiple video characteristics, each has a different resolution, frame rate, etc. We will shortly demonstrate the results for different resolutions of a video, and discuss the detected traffic violations.

We want to mention that we did not change the parameters of the fast or lane detectors at all, although the detector could be optimized for different resolutions. These parameters are set in away that is suitable for a resolution that is either 360p and 720p. However, we don't use any parameter tuning, and we test the robustness of the detector across different frame resolutions. For example, the search space of the lane marker detection area is the same for all resolutions. Hence, it performs poorly for the low resolution (144p) because it takes the most of the image in the search space, and for the high resolution (720p) because the search area is too small and it becomes hard to detect lane markers. The fix for this issue is to set the search area for the lane marker as a percentage of the image size. We found that this hard thresholding works very well, but we don't put the result due to the space limitations of the paper.

Another example for parameter tuning is the merge threshold of the fast detector. The merge threshold used to generate all the results is the same. However, it could be related to the size of the input image. We found that the detector struggles to find cars within the 144p video. Actually the images were not clear at all. Increasing the merge threshold could help in this case as multiple weak detections that are not accurate accumulate to select an object within the input image. Similarly, In the higher resolution, the merge search window size for the fast detector could be enlarged in order to capture cars with bigger sizes. These parameters can be optimally selected given the device type, the camera resolution, or the application settings.

An interesting test was performed when testing the fast detector using multiple resolutions of the same video [16], namely, 144p, 180p, 360p, and 720p. We generated these videos from the same YouTube link. We would like to mention that we trained the classifier with images that has different resolution. In Fig. 8(a), we can see that the number of fames is the same for all resolutions except for 144p, where some frame are lost during conversion. The figure shows that there is a pattern in the number of detections that follows the increase in the resolution. However, Fig. 8(b) shows that these numbers might be deceiving. For example, most of the detections do not result in the same pattern when detecting the motion, the lane markers, or relative positions. Hence, the fast detector should be trained for several resolutions in order to perform similarly regardless of the input video resolution. We did not use this criterion for the generation of data, but it is an interesting observation. Finally, we see a clear trend between the resolution and the processing time of the fast detector, and the lane-marker detector in Fig. 8(c). As the resolution increases, the processing time increases, and vice versa.

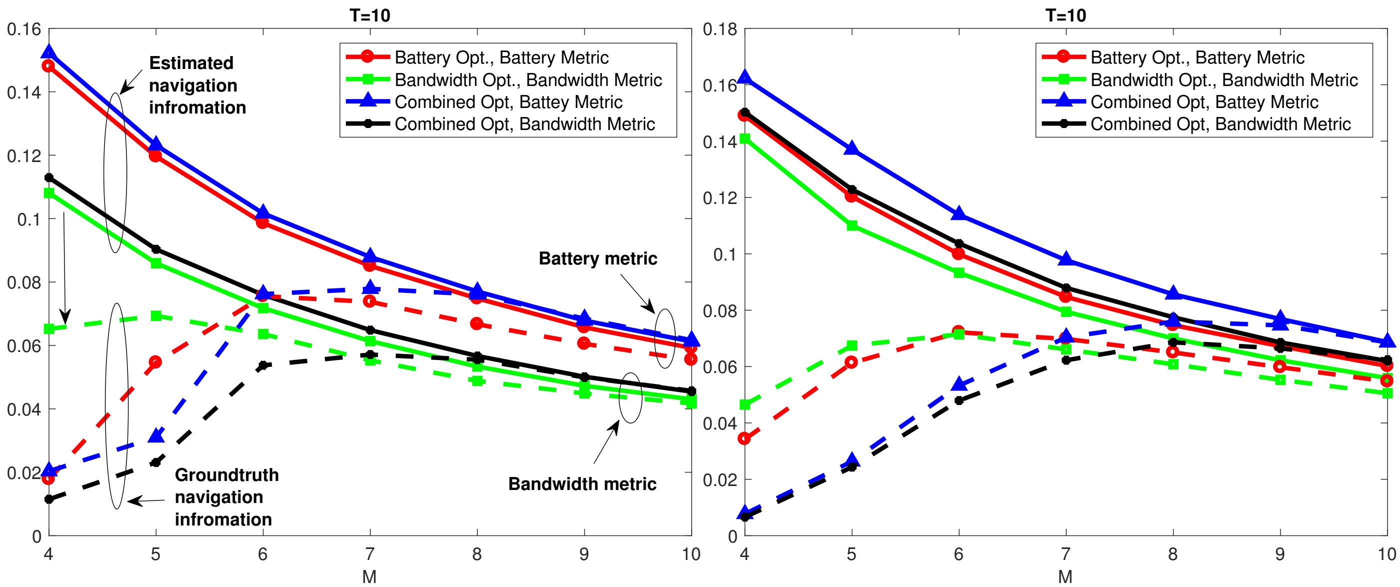
4) Battery and Bandwidth Optimization: For the evaluation of the optimization problem, we set the the optimization epoch time to $T = 10$. We then restrict the battery and bandwidth for each participant, by choosing a batter or bandwidth that are uniformly distributed in the range $[1, f(T, \delta)]$, where δ is an integer. We chose δ to enable smaller for the bandwidth constraints in the experiment (i.e., the bandwidth constraint are more stringent because some participants might not have unlimited data plans. However, the simulations can be performed in different ways as we show next. We set $B_i^Q = \lfloor \frac{T}{\delta} \rfloor$ for $\delta = 3$, and $D_i^Q = \lfloor \frac{T}{\delta} \rfloor$ for $\delta = 2$, and we plot the results in Fig. 6 (a). The figure shows the batter and bandwidth metrics vs the number of participants.

First, we observe that knowing the navigation information (dashed lines in the figure) has a lower use of the battery and bandwidth compared to the estimated navigation information (solid lines in the figure) regardless of the used metric or the optimization problem. The dashed lines are mostly higher than the solid lines, which means the scheme used more resources either in terms of batter or bandwidth. We also observe for this setting there is not a significant difference for performing the joint optimization versus a metric specific optimization (at least for the estimated navigation information case, i.e., solid lines). Hence, if the resources are scarce compared to the optimization length. It can also be inferred from the figure that as the resources increase, the resources utilization improves for the solid lines. However, when the navigation information is known, the combined battery and bandwidth optimization provides significant improvement over separate optimization, especially for the lower number of participants case.

We change the $B_i^Q = T - \delta$ for $\delta = 2$, and $D_i^Q = T - \delta$ for $\delta = 1$, and we plot the results in Fig. 6 (b). In this case, the bandwidth constrains should very close to the battery constraints as the difference in δ is small. As expected, the solid lines became closer to each others, which means that the solving either of the problems would not make a significant difference. Interestingly, losing the navigation information would also result in a significant gap between the solid lines and dashed lines, which means that using accurate navigation information (i.e., $b_i[t]$) would improve the utilization of resources for a small number of participants. However, if the number of participants is large, then, the navigation information cannot provide significant improvement.

V. Related Work

Car mobility and traffic violation detection has been studied from different perspectives. One part studies the computer vision interpretation of the car mobility. Another aspect is the communication methodology. Third, the different specific applications in the context. We provide the necessary review of the current works.



(a) (b)

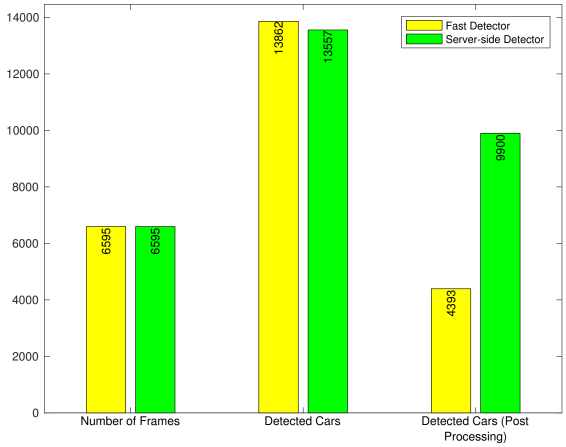
Fig. 6. Evaluation of the battery and bandwidth utilization metrics by solving the battery optimization problem (1)-(3), the bandwidth optimization problem (5)-(7), and the combined battery and bandwidth optimization problem (9)-(12).

A. Visual Processing of Traffic Information

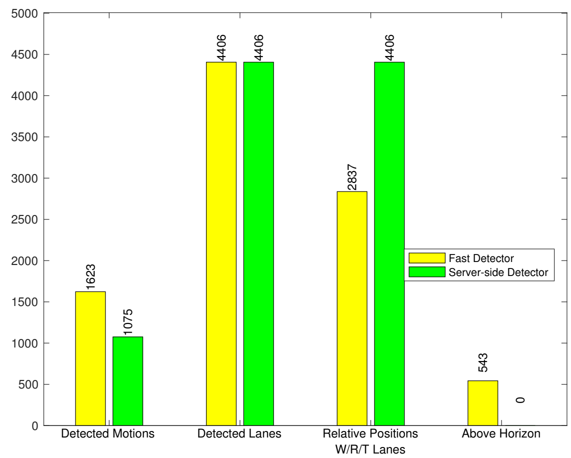
Haar-Like Cascade detector with AdaBoosting [33] is considered one of the fast detectors with a rate that can reach 50 frames per second because it deals with the integral image, and the learning procedure is performed through AdaBoosting which chooses a small number of features. In [34], a fast detector is used to classify the car from the background. In [23], the Cascade classifier is used similarly to alarm the driver if the car is closely tailgating the car in front of it [22]. The authors suggested some pre-processing of the data such as alignments of the cars in training set, and providing different version of the image in the original data set. Different other algorithms can be used for classifications and detection of cars i images. For example, in [35], a method combining temporal difference with and edge detector is used to detect cars. Such methods is widely proposed in the literature, and considered fast in terms of processing. However, these methods do not work properly with cars of different sizes (e.g., buses and trucks), and cause low accuracy of detection. Deep learning and neural networks resulted into exceptional performance in terms of localizing and detecting objects using traditional and Region-based convolutional neural networks (CNN and R-CNN) [30] [36] [37].

Lane detection has been a very active research topic in visual sensing of intelligent vehicles. The top-view or sometimes referred to as "bird's view" is a transformation method that is widely used in the literature for detecting the lanes on the road as in [38] [22] [39]. This method is formally referred to as, the inverse perspective mapping (IPM). In [22], the authors used the IPM procedure followed by a Hough transform. In [39], a robust and fast lane detector is proposed based on IPM, removing the outliers and reach excellent spline fitting using the well-known RANSAC algorithm. In [38], the authors used a combined IPM transformation and Hough transform line detector to detect the lanes. In [35], a modified Hough transform with a hardcoded search area is used for lane detection. A robust and fast line detector is proposed in [40] shows a potential to detect lines on roads.

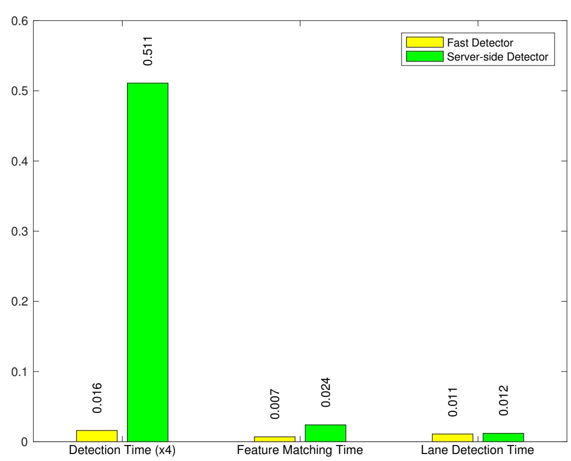
The distance between the vehicle ahead of the camera (or time to contact) has been studied in the literature. In [41], the authors proposed a robust, fast, and accurate estimation of time of contact between the car that has the camera within and the car in-front of it. In [22], the authors used the pinhole model to estimate the distance from the car ahead. In [42], a modified Harris corner detector is used to track vehicles motion. The famous feature detector used in the literature is Scale-Invariant Feature Transform (SIFT) [43]. A faster implementation version of SIFT is the Speeded-up Robust Features algorithm (SURF) [25]. It is also worth mentioning that combining other sensors with visual sensing could result in better understanding of the environment. A good source of information for visual processing on the roads is in [44]. The paper discusses many references and how they detect objects, lanes, motion using different machine learning algorithm, probabilistic methods, and using monocular and stereo-vision. A general architecture for video surveillance systems, namely, hierarchical and networked vehicle surveillance, is proposed where different used techniques in the literature are discussed [45]. The general architectural system is overviewed, but no performance evaluation is presented. A comparison between roadside (pole-mounted, stationary) and in-vehicle (mobile platforms) systems are presented in [46], but the authors focus on camera-based roadside monitoring systems, with special attention to omnidirectional setups. Another good source of information for safety analysis and the traffic behaviour at intersections with focus of visual sensing technology is in [47]. Different performance metrics are introduced, and are used to evaluate analyze some of the algorithms in the literature.



(a)

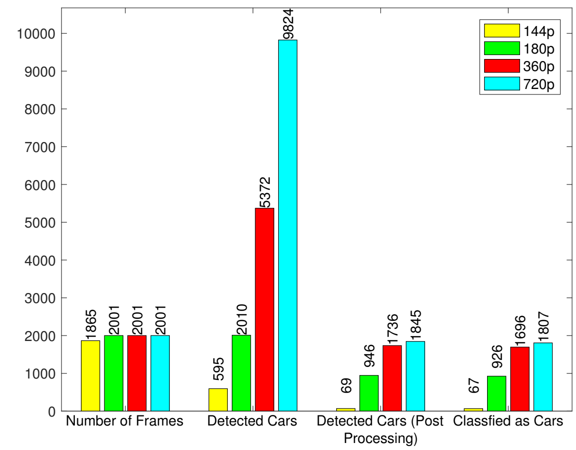


(b)

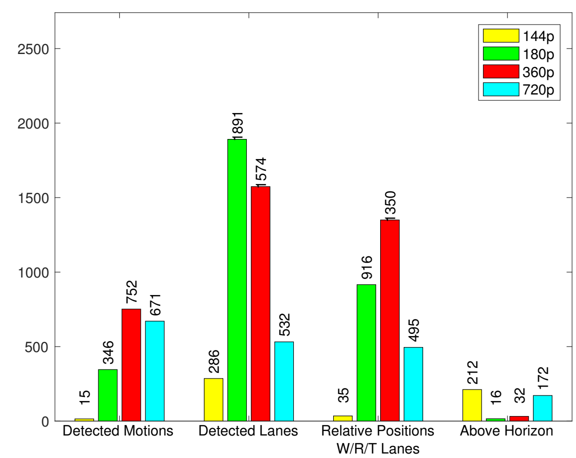


(c)

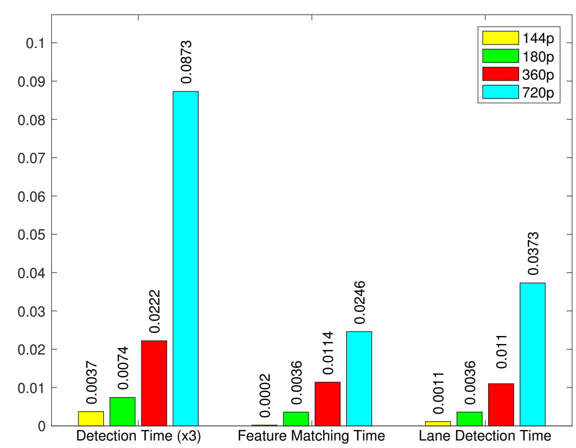
Fig. 7. Traffic violation metrics of the fast in-mobile and in-cloud detector using the testing video [16]. a) We compare the number of detected objects in the smartphone and the server, b) we illustrate the detected traffic violation within the smartphone and the server, and c) we compare the processing time for the different components of the system.



(a)



(b)



(c)

Fig. 8. Traffic violation metrics of the fast in-mobile and in-cloud detector using the testing video [16] with multiple resolutions. a) We compare the number of detected objects in the smartphone and the server, b) we illustrate the detected traffic violation within the smartphone and the server, and c) we compare the processing time for the different components of the system.

B. Communication Modes in Transportation Networks

Vehicles can communicate over the *dedicated short range communications* (DSRC) spectrum or the *cellular communication* infrastructure. The communication can occur in a regular cellular infrastructure mode, where the vehicles relay the information to the backend as a routing stage for the destination vehicle. Another option is that vehicles can directly communicate in a device-to-device mode using 5G or over the DSRC spectrum. A more delay-tolerant transmission can occur via WiFi offloading. Communication between vehicles has been widely studied in the recent literature [48], [49], [50]. In [51], the authors proposed a method for optimizing the collection of visual data crowd sourced by vehicular networks. In [52], a solution to mitigate the communication channel congestion is proposed using compressive sampling of packets. Reliability of transmission and delay of packets reception are still considered open research areas in vehicular communications.

C. Applications of Connected Visual Processing in Transportation

Normal cars on the roads can be used as active or passive participants for crowd sourcing. They can capture images or videos [53], [51], or provide the traffic state on the road. Moreover, visual understanding of the driving scene is becoming a crucial part of autonomous vehicles [54]. Connected vehicles can be used to optimize vehicles routing and fuel consumption [55], or optimize traffic light operation [56]. Visual information of the roads can feed the algorithms with additional information of the road state, and the number of vehicles.

Safety is the top issue in driving, and visual processing can be an asset to complement the communication capability where the missed information can be crucial. Unlike most of the work that is currently focused on autonomous driving, a costly equipment that can be given to specific users such as transportation officers, we propose a unique solution that is low-cost, can be used for a variety of driving detection, and is pervasive in nature.

VI. Conclusion

This paper propose a novel solution that provides accurate and low-cost traffic violation detection. We design a framework that uses the current cameras in smartphones, and implemented a fast detector that works on the smartphones. The fast detector works with good accuracy. Therefore, we propose the use the deep CNN classifier to distinguish cars from backgrounds, which has a very high accuracy. After that, we process the videos on the servers-side using a more accurate detector that tolerate-delay. After detecting cars in different frames, we use feature matching following by an outlier detector algorithm to match the positions of the cars within the frames. The motion detector can classify the motion of the vehicle in different directions and relative to the lane markers. The sensory information of the phone such as GPS, gyroscope, compass, and accelerometer are used to detect the relative behaviour to the monitoring car. This can be an over- or under-speeding vehicle, a vehicle driving over the lane marker or switching lanes, etc. We demonstrated that the accuracy of the proposed system is directly related to the accuracy of the used computer vision algorithms. Moreover, the more data are used to train object detectors and classifier, the better becomes the accuracy. We also proposed a method to utilize the resources of the participants according to the limited battery and bandwidth of each participant, and the availability of the navigation information. We show that the navigation information is important for a lower number of participants, but loses importance as the number of participants increases. We also demonstrated that battery and bandwidth optimization can be combined, but that does not provide significant improvement unless the amount of resources have a large variance (i.e., the battery and bandwidth are not close in value on average). Finally, we conclude that our solution can be pervasive and can be a low-cost asset to any driving detection facility, and that we do not compete with self-driving vehicles, rather, the proposed solution can be integrated to them.

Acknowledgment

This author would like to thank the Deanship of Scientific Research at Umm Al-Qura University for the financial support of the project #17-COM-1-02-0002.

References

[1] S. Kaewunruen, J. M. Sussman, and A. Matsumoto, "Grand challenges in transportation and transit systems," *Frontiers in Built Environment*, vol. 2, p. 4, 2016. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fbuil.2016.00004

[2] "Traffic accidents occur every minute in ksa," http://www.arabnews.com/saudi-arabia/news/717596, accessed: 2019-03-15.

[3] G. Karagiannis, O. Altintas, E. Ekici, G. Heijenk, B. Jarupan, K. Lin, and T. Weil, "Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions," *IEEE communications surveys & tutorials*, vol. 13, no. 4, pp. 584–616, 2011.

[4] "Mobileye," http://www.mobileye.com/en-us/, accessed: 2019-03-01.

[5] "everdrive," https://www.everquote.com/everdrive/, accessed: 2019-05-01.

[6] "Caltech car dataset," http://www.vision.caltech.edu/html-files/archive.html, accessed: 2019-11-01.

[7] "Caltech car dataset – extended," http://www.robots.ox.ac.uk/~vgg/data/data-cats.html, accessed: 2019-11-01.

[8] "Caltech background dataset – extended," http://www.robots.ox.ac.uk/~vgg/data/data-cats.html, accessed: 2019-11-01.

[9] "Caltech road dataset," http://vasc.ri.cmu.edu/idb/images/road/, accessed: 2019-11-01.

[10] "Cars dataset," http://ai.stanford.edu/~jkrause/cars/car_dataset.html, accessed: 2019-11-01.

[11] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[12] "Karlsruhe dataset: Labeled objects (cars + pedestrians)," http://www.cvlibs.net/datasets/karlsruhe_objects/, accessed: 2019-11-01.

[13] A. Geiger, C. Wojek, and R. Urtasun, "Joint 3d estimation of objects and scene layout," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[14] "Youtube," https://www.youtube.com/watch?v=znlKpb2TDe8, accessed: 2019-11-01.

[15] "Youtube," https://www.youtube.com/watch?v=J_AHVG7mutA, accessed: 2019-11-01.

[16] "Youtube," https://www.youtube.com/watch?v=D5WxoZJsZpM, accessed: 2019-11-01.

[17] "Youtube," https://www.youtube.com/watch?v=9G4dGTy5Cyw, accessed: 2019-11-01.

[18] "Youtube," https://www.youtube.com/watch?v=p5T8RzYE60M, accessed: 2019-11-01.

[19] "Youtube," https://www.youtube.com/watch?v=hHfoLnCN2no, accessed: 2019-11-01.

[20] "Youtube," https://www.youtube.com/watch?v=xkvKV5wsYUE, accessed: 2019-11-01.

[21] "Youtube," https://www.youtube.com/watch?v=GceaIO5OmiU, accessed: 2019-11-01.

[22] C.-W. You, N. D. Lane, F. Chen, R. Wang, Z. Chen, T. J. Bao, M. Montes-de Oca, Y. Cheng, M. Lin, L. Torresani, and A. T. Campbell, "Carsafe app: Alerting drowsy and distracted drivers using dual cameras on smartphones," in *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '13. New York, NY, USA: ACM, 2013, pp. 13–26. [Online]. Available: http://doi.acm.org/10.1145/2462456.2465428

[23] T. Wang, G. Cardone, A. Corradi, L. Torresani, and A. T. Campbell, "Walksafe: A pedestrian safety app for mobile phone users who walk and talk while crossing roads," in *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*, ser. HotMobile '12. New York, NY, USA: ACM, 2012, pp. 5:1–5:6. [Online]. Available: http://doi.acm.org/10.1145/2162081.2162089

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[25] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[26] P. H. Torr and A. Zisserman, "Mlesac: A new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 138–156, 2000.

[27] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Discriminatively trained deformable part models, release 5," http://people.cs.uchicago.edu/ rbg/latent-release5/.

[28] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[29] "clarifai," https://clarifai.com, accessed: 2017-03-01.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[31] "Arada systems," http://www.aradasystems.com, accessed: 2017-04-01.

[32] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[33] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[34] K. Chang, B. H. Oh, and K. S. Hong, "An implementation of smartphone-based driver assistance system using front and rear camera," in *2014 IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2014, pp. 280–281.

[35] S. J. W. Tang, K. Y. Ng, B. H. Khoo, and J. Parkkinen, "Real-time lane detection and rear-end collision warning system on a mobile computing platform," in *2015 IEEE 39th Annual Computer Software and Applications Conference*, vol. 2, July 2015, pp. 563–568.

[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[37] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[38] C.-Y. Fang, W.-H. Hsu, C.-W. Ma, and S.-W. Chen, "A vision-based safety driver assistance system for motorcycles on a smartphone," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, Oct 2014, pp. 328–333.

[39] M. Aly, "Real time detection of lane markers in urban streets," in *2008 IEEE Intelligent Vehicles Symposium*, June 2008, pp. 7–12.

[40] R. Grompone von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "On straight line segment detection," *Journal of Mathematical Imaging and Vision*, vol. 32, no. 3, p. 313, 2008. [Online]. Available: http://dx.doi.org/10.1007/s10851-008-0102-5

[41] B. K. P. Horn, Y. Fang, and I. Masaki, "Time to contact relative to a planar surface," in *2007 IEEE Intelligent Vehicles Symposium*, June 2007, pp. 68–74.

[42] Z. Di and D. He, "Forward collision warning system based on vehicle detection and tracking," in *2016 International Conference on Optoelectronics and Image Processing (ICOIP)*, June 2016, pp. 10–14.

[43] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[44] S. Sivaraman and M. M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 4, pp. 1773–1795, Dec 2013.

[45] B. Tian, B. T. Morris, M. Tang, Y. Liu, Y. Yao, C. Gou, D. Shen, and S. Tang, "Hierarchical and networked vehicle surveillance in its: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 557–580, April 2015.

[46] S. R. E. Datondji, Y. Dupuis, P. Subirats, and P. Vasseur, "A survey of vision-based traffic monitoring of road intersections," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 10, pp. 2681–2698, Oct 2016.

[47] M. S. Shirazi and B. T. Morris, "Looking at intersections: A survey of intersection monitoring, behavior and safety analysis of recent studies," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, pp. 4–24, Jan 2017.

[48] W. Chen, *Vehicular communications and networks: Architectures, protocols, operation and deployment*. Elsevier, 2015.

[49] C. Campolo and A. Molinaro, "Multichannel communications in vehicular ad hoc networks: a survey," *IEEE Communications Magazine*, vol. 51, no. 5, pp. 158–169, 2013.

[50] W. Viriyasitavat, M. Boban, H.-M. Tsai, and A. Vasilakos, "Vehicular communications: Survey and challenges of channel and propagation models," *IEEE Vehicular Technology Magazine*, vol. 10, no. 2, pp. 55–66, 2015.

[51] W. Alasmary, H. Sadeghi, and S. Valaee, "Crowdsensing in vehicular sensor networks with limited channel capacity," in *2013 IEEE International Conference on Communications (ICC)*, June 2013, pp. 1833–1838.

[52] W. Alasmary and S. Valaee, "Compressive sensing based vehicle information recovery in vehicular networks," in *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)*, July 2013, pp. 700–705.

[53] M. Gerla, J.-T. Weng, and G. Pau, "Pics-on-wheels: Photo surveillance in the vehicular cloud," in *Computing, Networking and Communications (ICNC), 2013 International Conference on*. IEEE, 2013, pp. 1123–1127.

[54] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "Dr(eye)ve: A dataset for attention-based tasks with applications to autonomous and assisted driving," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2016, pp. 54–60.

[55] M. Alsabaan, W. Alasmary, A. Albasir, and K. Naik, "Vehicular networks for a greener environment: A survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1372–1388, Third 2013.

[56] S. El-Tantawy, B. Abdulhai, and H. Abdelgawad, "Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (marlin-atsc): Methodology and large-scale application on downtown toronto," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1140–1150, Sept 2013.

HarmonyMoves: A Unified Prediction Approach for Moving Object Future Path

Mohammed Abdalla[1], Hoda M. O. Mokhtar[2], Neveen ElGamal[3]
Faculty of Computers and Artificial Intelligence
Beni-Souef University[1]
Cairo University[2,3]

***Abstract*—Trajectory prediction plays a critical role on many location-based services such as proximity-based marketing, routing services, and traffic management. The vast majority of existing trajectory prediction techniques utilize the object's motion history to predict the future path(s). In addition to, their assumptions that the objects' moving with recognized patterns or know their routes. However, these techniques fail when the history is unavailable. Also, these techniques fail to predict the path when the query moving objects lost their ways or moving with abnormal patterns. This paper introduces a system named *HarmonyMoves* to predict the future paths of moving objects on road networks without relying on their past trajectories. The system checks the harmony between the query object and other moving objects, after that if the harmony exists, this means that there are other objects in space moving like the query object. Then, a Markov Model is adopted to analyze this set of similar motion patterns and generate the next potential road segments of the object with their probabilities. If the harmony does not exist, *HarmonyMoves* considers this query object as abnormal object (*object lost the way and needs support to return back known routes*), for this purpose *HarmonyMoves* employed a new module to handle this case. A fundamental aspect of *HarmonyMoves* lies in achieving a high accurate prediction while performing efficiently to return query answers.**

***Keywords*—Trajectory prediction; machine learning; moving objects**

I. Introduction

Location-based services that consider the future paths and locations of moving objects proved to be essential in several daily life activities. Uses of location-based services includes proximity-based marketing; in which companies push ads notifications only to customers within the same geographic region, travel information; in which the user can be provided by real-time information, such as traffic updates or weather reports and plan trips accordingly, traffic management; in which drivers can predict the congested traffic regions.

Several techniques have been proposed to predict the possible future paths and destinations of moving objects and explores their results in better understanding of the human mobility into location-based services [3], [9], [10], [24], [26]. Overall, most of the existing prediction techniques depend basically on the trajectories of the moving objects' saved on the system beforehand to be able to predict the future path and destinations. Nevertheless, these techniques suffer from one or more weaknesses. These techniques (1) fail to predict future paths and destinations when the query object's history does not exist during the model training, (2) some of these techniques consider assumptions like moving objects following the shortest path or following preferred routes to reach the needed destination or moving in linear movements, which usually fails and turns to be non logical in many scenarios, (3) some of these techniques don't consider the objects that lost their ways or moving with unrecognized motion patterns, (4) suffer from efficiency and technical limitations.

This paper introduces a novel system named *HarmonyMoves* proposed to provide an efficient future path prediction in case that the query object's historical motions are absent, in addition to guiding the query objects that lost their ways to follow the correct paths and reach to their destinations in a safely manner.

The idea of *HarmonyMoves* is to identify if there are other moving objects currently in space moving like the query object or not. If the system finds similar objects moving like the query object, the system considers the query object as normal object, and trained for similar objects trajectories to anticipate the possible future path of the query object. In case that the system identifies that there are no other objects moving like the query object, the system considers the query object as anomalous , and starts to guide the query object in terms of the nearest moving object in vicinity.

Contributions. The main contributions of this research are the following:

1) This work addresses the prediction of moving objects' future path in case that their past trajectories are absent.
2) This work introduces a novel similarity function to help all types of moving objects to predict their routes and find their final destinations, in addition to, in case of normal objects this function leads to a significant increase in prediction efficiency by removing irregularities in the input data and make the prediction model trains on data similar to each other.
3) We devise a novel algorithm to handle moving objects that lost their paths and need to predict future path or reach final destinations.

The rest of this paper is organized as follows. Section 2 studies related work and investigates different directions in the area of trajectory prediction. Section 3 formally defines the problem. The architecture of the *HarmonyMoves* system along with its basic components is illustrated in Section 4. Section 5 experimentally evaluates *HarmonyMoves*. Finally, Section 6 concludes the paper.

II. Related Work

This section surveys the previous studies which predict the future paths of the moving objects. Overall, this section covers two major related directions; namely path prediction and predictive search queries.

A. Path Prediction

Different studies have been introduced to predict future paths of moving objects [2], [5], [6], [9], [17], [21], [27]–[30]. However, some of these studies predict the complete paths [5], [6], [27], others predict partial paths [2], [17], [28], [29], whereas some predict the final destination [9], [21], [30].

For predicting the complete path of the end-users, in [6], a similarity trip algorithm is developed to predict the end-to-end route of a vehicle based on vehicle's past trips observations. The proposed algorithm matches the first part of a driver's current trip with one of the set of previously observed trips. In [27], authors develop a driving path prediction technique that employed Hidden Markov Model (HMM). This technique predicts accurately a vehicle's entire path as early in a trip's lifetime as possible without knowing origins and destinations in advance. Authors in [5] propose a novel algorithm for predicting a driver's route based on a probabilistic prediction of the driver's destination. The algorithm is based on only one parameter that shows how efficiently a driver drives. When this parameter is computed, the algorithm does not need to store a history of drivers' trips, and it works in places a driver has never visited.

For predicting the partial path of the end-users, in [29], authors propose a technique which adopted Hidden Markov Model to predict future road segments from the complete driving path. The authors in their solution neglects the traffic conditions. In contrast, authors in [17], [28] take into their account traffic conditions that change dynamically and prove that the work introduced in [29] is not adequate to capture variable order Markov dependencies. Some studies integrate semantic information of routes to improve prediction. Authors in [2] propose an approach to improve partial route prediction by considering semantic information associated with routes such as day and time of departure.

For predicting the finial destinations of the end-users, in [9], [30], prediction models are proposed to detect the trajectory patterns that occurred frequently and utilize these patterns to predict the most potential destination of the moving objects. Authors in [21] introduce a technique predict the final destination of vehicle trips based on their initial partial trajectories. The technique starts by clustering of trajectories that describes user behaviour. Then, it models main traffic flow patterns by a mixture of 2d Gaussian distributions.

B. Predictive Queries

According to the study in [19], a predictive search query is conducted as follows: "*a query that retrieves the set of moving objects' which will intersect a query window during a future time interval*". Several studies have been in this direction [4], [7], [16], [19], [20], [25].

In [7], *Panda* system is presented, *Panda* aims to support long-term query prediction. In [16], a novel index structure, named Predictive tree (P-tree) is proposed for processing predictive queries against moving objects on road networks. Efficient techniques are introduced in [4], [20], [25] that try to optimize the performance of predictive window queries. In [1], the authors introduce the iRoad framework for evaluating predictive queries on moving objects for road networks. In [8], authors review the current research trends and present their related applications in the field of predictive spatiotemporal queries processing.

TABLE I. Moving Objects Trajectories

Trajectory	Road Segments
τ_1	e_6, e_7, e_{16}, e_{14}
τ_2	e_8, e_{16}, e_{14}
τ_3	e_7, e_{16}, e_{14}
τ_4	e_7, e_{16}, e_{19}
τ_5	e_7, e_{16}, e_{17}
τ_6	e_8, e_{16}, e_{17}
τ_7	e_{11}, e_9, e_7, e_{16}
τ_8	e_{18}, e_{13}, e_{12}

HarmonyMoves distinguishes itself from the above researches in the following: (1) *HarmonyMoves* is considered as a novel learning model which performs the prediction without any prior knowledge of the query object's motions history, (2) *HarmonyMoves* proves efficient and accurate results by providing novel similarity approach which neglects irrelevant input features from prediction process, (3) *HarmonyMoves* is considered as first approach that handles moving objects which loses their ways, and (4) *HarmonyMoves* produces accurate results compared with other state-of-art methods.

III. Preliminaries

This section explores the main terminologies that will be used through the rest of the paper and then highlights the problem definition.

A. Definitions

Definition 1: Road network graph G(N, E) consists of a set of nodes N that represent road intersections, and a set of edges E that represent road segments.

Fig. 1 presents an illustrative example of a road network graph. Table I contains a sample of objects' trajectories τ_1 to τ_{14} moving over the given road network graph(G).

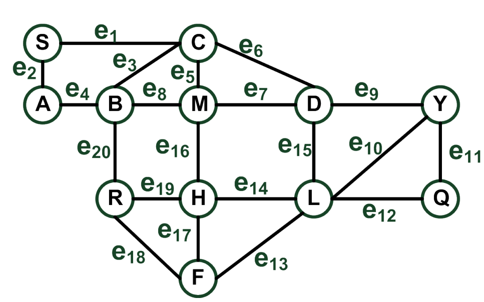


Fig. 1. Road Network Graph

Definition 2: Trajectory τ is an ordered sequence of edges travelled by a moving object, i.e., τ={e_1, e_2, e_3,....e_n}

Definition 3: Future road segments $\mathcal{F}$ is defined as how many future road segments query object aims to predict.

Definition 4: Harmony it represents the total *Similarity* value between the query object's trajectory Q_τ and another moving object's trajectory τ_i. The equation 1 presents the harmony computing. The Harmony value is computed by getting the maximum length between the query object's trajectory Q_τ and another moving object's trajectory τ_i, then computes the edit distance between (Q_τ,τ_i). After that the result of edit distance is subtracted from maximum length between the query object's trajectory Q_τ and another moving object's trajectory τ_i. Finally, the value subtracted is divided by maximum length between the query object's trajectory Q_τ and another moving object's trajectory τ_i. The edit distance is computed based on Levenshtein distance metric [15], this metric measures the difference between two sequences. It computes the total number of edits (insertions, deletions or substitutions) needed to change one sequence into the other.

$$Harmony = \frac{MaxLength(Q_\tau, \tau_i) - E_{Distance}(Q_\tau, \tau_i)}{MaxLength(Q_\tau, \tau_i)} \tag{1}$$

Definition 5: Prediction confidence $\mathcal{P}$ is a probability value that reveals a certainty degree which the query object will travel through the predicted path. The higher the value, the more accurate the prediction is.

Definition 6. Number of Previously Travelled Segments (Prediction Order) λ reveals to the last number of segments moved by query object's trajectory over the road network. For example: assume $\tau_i = \{e_1, e_2, e_3, e_4\}$ and λ =2, this means that the extracted segments will be (e_3, e_4).

B. Problem Statement

Assume a query object's trajectory Q_τ moving in its current trip and a set of another objects $Trajectories_{Current}$ moving in their current trips. Let Q_τ needs to predict the future path, in case that its own historical movements data is unavailable or inaccessible. In addition to, Q_τ needs to predict the future path when it moves with unknown motion patterns or lost their ways.

The aim of this research is to accurately predict the future paths of the query object and perform efficiently by returning the query responses in a reasonable time.

The significant of this research lies in solving critical cases such as; (1) when a query object does not own previous trips to be used for prediction purposes or the query object is not allowed to use the his past trips for security concerns, and (2) when a query object lost his way or moving with unknown motion patterns and needs to predict the future path while there is no other objects moving like him.

IV. System Architecture

This section presents the proposed approach (*HarmonyMoves*) for predicting the future path of the query object.

Main Idea. The main idea of *HarmonyMoves* is to anticipate the possible future paths or destinations and returns them to the query object. The proposed solution starts by measuring the similarity, which in this work named as Harmony, between the query object and other objects moving currently in the same road network.Next, the system checks if there are other objects moving like the query object (*Harmony¿0*), the system contains a module named *"normal objects handler module"* to deal with this case, this module starts by getting the similar trajectories, after that utilizes these trajectories to generate predictions by using a learning model. It is important to note that predictions generated is a map (key/value) pairs, which composed of travelled road segments as a key and possible future segments as a value. Finally, after the predictions generated, the system query the created map by the last segments moved by the query object and returns the possible future segments. Next, if the system checks the harmony and discovers that there are no other objects moving like the query object (*Harmony=0*), the system considers the query object as a vagrant moving object. Vagrant objects represent the objects that lost their ways or objects moving with unknown roads needs to reach their destinations or predict next segments to reach to the known routes. More specifically, the system contains a module named *"vagrant objects handler module"* to deal with this type of objects, first the system the system tracks the nearest moving objects and based on his trajectory, the system suggests a possible route to reach the destination or suggests multi next road segments.

Fig. 3 describes the architecture of the proposed solution *HarmonyMoves* which composed of three main components namely, the Harmony Checker Module, Normal Objects Handler Module, and Vagrant Objects Handler Module. *HarmonyMoves* receives as input the moving objects' trajectories and the query object's trajectory, and returns the query object's future path or destination.

Algorithm. Algorithm **??** illustrates the pseudo code of *HarmonyMoves*. The algorithm receives, (a) the current query object's trajectory, (b) other moving objects' trajectories currently in the road network, (c) prediction order, and (d) a number of future road segments required to be predicted. As output, the algorithm returns the query object's predicted path or possible destination. The algorithm has three main steps that are briefly described as follows:

Step1: Harmony checker. This step checks the similarity between the query object and other moving objects, in line 4 the algorithm checks the harmony between the query object and other objects currently moving with the query object in the road network. The proposed module that accomplishes this job is explained later in Section IV-A.

Step2: Handling Normal Query Objects. The objective of this step is to handle the prediction process in case that the query object has similar objects currently moving like it currently in the system. The proposed module that accomplishes this job is explained later in Section IV-B.

Step3: Handling Vagrant Query Objects. The objective of this step is to handle the prediction process in case that the query object has no similar objects currently moving like it currently in the system. The proposed module that accomplishes this job is explained later in section IV-C.

A. Harmony Checker Module

Main Idea. The main idea of the Harmony Checker Module is to extract trajectories set of objects that currently move same as the query object. Then, the module checks if this set of trajectories is empty, the *HarmonyMoves* considers the query object as vagrant object, otherwise it considers it as normal one. For this purpose, the harmony checker module calculates the harmony score between the query object's trajectory and all other trajectories of objects that are currently moving in the space by utilizing proposed equation 1. If the Harmony value is equal to 0, the Harmony Checker Module considers the query object is vagrant, otherwise it considers it as normal one.

Example. Assume that the query object's trajectory is τ_7, and assume that the other objects currently moving in road network are τ_2 e_8, e_{16}, e_{14}, and τ_3 e_7, e_{16}, e_{14}. The module starts by calculating the harmony between the query object τ_7 and τ_2,τ_3. For τ_2, the module first gets the max length between τ_7 and τ_2 which is 3, then computes the edit distance between τ_7 and τ_2 which is 1. As a result, the module subtracted the edit distance score from max length between τ_7 and τ_2 divided by the same length, so the final result is Harmony $= \frac{1}{3}$. Similarly, the system computes the harmony between τ_7 and τ_3 which is Harmony $= \frac{2}{3}$. Finally, the system considers τ_7 as normal query object. For second case, assume that the query object τ_8 and other moving objects are τ_7 and τ_2, the harmony in all case will be equal to 0 and the system considers τ_7 as vagrant object.

B. Normal Objects Handler Module

This section presents a new module that handles the cases when the query object have other objects moving similar like it in the system. This module builds a predictions model from the list of similar trajectories and, then returns the query object's predicted path. The module composed of two main sub modules namely, Predictions Builder Module, and Path Prediction Module.

1) Predictions Builder Module: Main Idea. The input to this module is the set of similar trajectories that are moving same as the query object. The output of this module is a hash-map. This map composed of Key and value pairs, where the key is an ordered set of previous segments (already-visited) in a trajectory and the value is a sequence of future (want to visit) segments and its probability $\mathcal{P}$. The prediction function employed in this module is developed based on notions of Hidden Markov Model(HMM) [14]

Hidden Markov Model (HMM). The Markov Model represents the sequence of moved segments as a sequence of *X(i)*, where i is the offset of the segemnt in the order they are encountered inside the trajectory. As stated in [11], HMM refers to the trajectory edges as ..., X(-2), X(-1), X(0), X(1), X(2),..., where X(0) is the object's current road edge. X(-1) and X(-2) refer to one and two previous steps, respectively. X(1) and X(2) refer to the unknown future road segments to be predicted. HMM produces a probabilistic prediction over future road segments based on the past moved segments. For example, $\mathcal{P}$[X(1)] represents the probability of a one segment ahead of the object's current location. $\mathcal{P}$[X(2)] is the probability of two segments ahead, and so on. The first order Markov model says that the probability $\mathcal{P}$[X(1)] for the next road segment is independent of all the object's previous history except for X(0), the current segment:

$$\mathcal{P}[X(1)|X(0), X(-1), X(-2), ...] = \mathcal{P}[X(1)|X(0)] \quad (2)$$

Similarly, a second order Markov model will depend on the two most recent edges,$\mathcal{P}$[X(1)|X(-1),X(0)]. Overall, a n^{th} order Markov model, ($n \geq 1$), can be created to anticipate the m^{th} future edge ($m \geq 1$). The general n^{th} order model can be expressed as following:

$$\mathcal{P}_n[X(m)] = \mathcal{P}[X(m)|X(-n+1), X(-n+2), ..., X(0)] \quad (3)$$

Fig. 2 descries the results produced by this module.

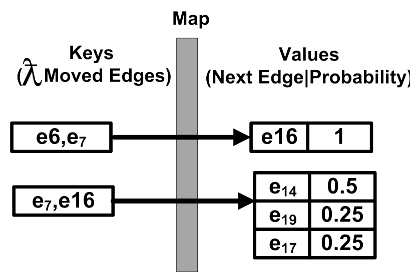


Fig. 2. Predictions Builder Module Results

2) Path Prediction Module: Main Idea. The main idea of this module is to return the possible future segments to the query object. First, the module detects the prediction order configured, then based on the value of prediction order, the module gets the last segments moved by the query object and considers it as key. After that, the module query the created map by the key and return the future path.

Example. Let's the query object is τ_7, assume the value configured for prediction order λ=2, and the number of future steps needs to be predicted $\mathcal{F}$ is 1. According to this setup, the module constructed the key as (e7,e16) and then query the map created as per Figure 2. The path prediction Module returns next segments to the query object, first, $\{e_{14}$,$\mathcal{P} = 0.5\}$,$\{e_{19}$,$\mathcal{P} = 0.25\}$, and $\{e_{17}$,$\mathcal{P} = 0.5\}$

C. Vagrant Objects Handler Module

This section presents a new module that handles the cases when the query object has no similar objects moving like it in the system.

Main Idea. The main idea behind this module is to handle the prediction process in case that there is no similar trajectories currently moved in the system anymore and the query object issue a query asking for the future path. The paper named these objects as Vagrant Moving objects. Vagrant objects represent the objects that lost their ways or objects moving with unknown roads need to reach their destinations. Handling these objects is significant as most of road accidents

Algorithm 1 *HarmonyMoves:* Algorithm

INPUT: Query object's trajectory Q_τ, Moving objects' current trajectories $Trajectories_{Current}$,Future steps $\mathcal{F}$
SET $SimilarTrajectoriesList\ S \leftarrow \phi$
/* Step 1: compute harmony */
S = Compute_Harmony(Q_τ,$Trajectories_{Current}$)
/* Step 2: checks Q_τ is vagrant object or normal one */
if $S \neq \phi$ **then**
    Call Normal Objects Handler Module
else
    Call Vagrant Objects Handler Module
end if
OUTPUT: Return $QueryResult$

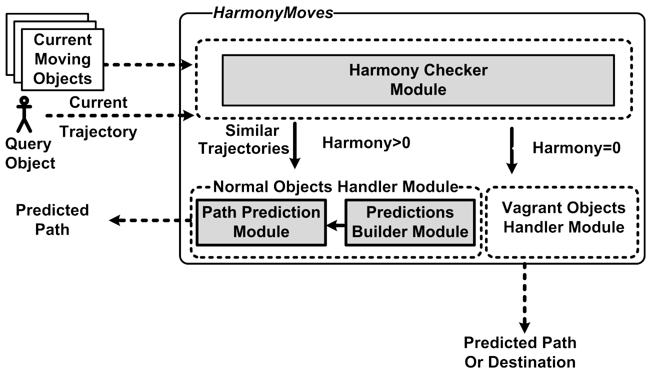


Fig. 3. *HarmonyMoves* Architecture

Algorithm 2 Unknown Destinations Handler

1: **procedure** GET_FUTURE_PATH
2: **INPUT:** Query object's trajectory Q_τ, $Future\ Steps\ \mathcal{F}$, Moving object's trajectories $Trajectories_{Current}$, Distance d
3: **OUTPUT:** Return $Predicted\_Path$
4:     /* Set compared Trajectory */
5:     **SET** $Compared_\tau \leftarrow \phi$
6:     /* Last edge in query object trajectory */
7:     **SET** $Q(CurrentEdge)_\tau \leftarrow \phi$
8:     /* edges connected direct to query object current node */
9:     **SET** $edgesList \leftarrow \phi$
10:     /* All expected paths */
11:     **SET** $PathList \leftarrow \phi$
12:     **if** $\mathcal{F} = 0$ **then**
13:         *Exit and return* $Predicted\_Path$
14:     **end if**
15:     **for** *i=0 to* $\mathcal{F}$ **do**
16:         $Compared_\tau$ = call R-Tree to get nearest Trajectory to Q_τ from $Trajectories_{Current}$ within d
17:         $Q(CurrentEdge)_\tau$ = Current edge of Q_τ
18:         *Put in edgesList edges connected by* $Q(CurrentEdge)_\tau$
19:         **for** ***each*** *edge* e_i *in* $edgesList$ **do**
20:             *FetchedPath=* Q_τ + e_i
21:             *Add* $FetchedPath$ *to* $PathList$
22:         **end for**
23:         **for** ***each*** $FetchedPath$ *in* $PathList$ **do**
24:             *ComputeSimilarity(*$FetchedPath$*,*$Compared_\tau$*)*
25:             *Set Most Similar FetchedPath as* Q_τ
26:             *Set Most Similar FetchedPath as* $Predicted\_Path$
27:         **end for**
28:     **end for**
29:     *Get_Future_Path(*$Q_\tau, \mathcal{F}-1, Trajectories_{Current}$*)*
30: **end procedure**

and people deaths coming from similar cases. This module handles 2 possible scenarios; (1) vagrant objects don't know their destinations and need to update their paths per time, and (2) vagrant objects know their destinations and don't know how to reach. For unknown destinations, the module starts by using R-Tree [22] to get the nearest object's trajectory within a specific distance. After that, the module considers this trajectory as a compared trajectory. Next, the module identifies the current edge of the query object, fetches the next connected edges to this current edge and creates paths. The module iterates over these created paths and computes the similarity between each iterated path and the compared trajectory by using the Hausdorff [23] algorithm. As a result, the module considers the most similar path to the compared trajectory as new query object trajectory and make recursive calls until predicting all the required future steps. For known destinations, the module starts by using R-Tree to get the nearest object's trajectory within a specific distance. After that, the module considers this trajectory as a compared trajectory. Next, the module identifies the current edge of the query object, fetches the next connected edges to this current edge and create paths. The module iterates over these created paths and computes the similarity between each iterated path and the compared trajectory by using the Hausdorff algorithm. As a result, the module considers the most similar path to the compared trajectory as new query object trajectory and make recursive calls until the query object reach to its destination.

1) Unknown Destinations: This section presents the algorithm and illustrative example for case 1; vagrant objects don't know their destinations and need to update their paths per time.

Algorithm. Algorithm 2 illustrates the pseudo code of the proposed Handler Module that handle case 1 defined before for vagrant objects. The algorithm takes four input parameters, (a) Query object's trajectory Q_τ, (b) $Future\ Steps\ \mathcal{F}$, (c) Moving object's trajectories $Trajectories_{Current}$, and (d)Distance d. In line 15, the algorithm starts by iterating over the number of future steps $\mathcal{F}$. In each iteration, the algorithm gets the compared trajectory from the objects currently moved with the query object by using the R-tree within specific distance, line 16. In line 17, the algorithm identifies the current edge of the query object, obtains the edges directly connected to the current node and saved them in $edgesList$. Then, the algorithm iterates over the $edgesList$, and for each iterated edge, the algorithm append this edge to the query object trajectory, then consider this as new path. Each path is saved to $PathList$ and algorithm considers it as fetched path. After that, in line 23, the algorithm iterates over the $PathList$, and computes the similarity between each path and the compared trajectory. Hence, the algorithm makes the most similar path as the predicted path. Besides this, the algorithm considers it as a new query object's trajectory, (lines 23-27). Then, algorithm decrease the number of future steps by 1 and make a recursive calls each time by the new calculated future steps, new created query object's trajectory. Finally, The algorithm check if the $\mathcal{F} = 0$, it skips and returns back the predicted path (Lines 12-14).

Example. Assume that the query object's trajectory is $\tau_7(e_{11}, e_9, e_7, e_{16})$, assume the number of future steps to be predicted ($\mathcal{F}$) is 2, assume that $SimilarTrajectoriesList$ S is ϕ, and assume d is 1 Kilo Meter. First, the algorithm gets the compared trajectory $Compared1_\tau$, say for example (e_3, e_1, e_2). Second, gets the connected edges to the current node of the query object which are e_{14}, e_{19}, e_{17}. Accordingly, the algorithm creates 3 paths: $path_1\{e_{11}, e_9, e_7, e_{16}, e_{14}\}$,$path_2\{e_{11}, e_9, e_7, e_{16}, e_{19}\}$, and $path_3\{e_{11}, e_9, e_7, e_{16}, e_{17}\}$. After that, the algorithm computes the similarity between each path and $Compared1_\tau$, as result algorithm returns $path_2$ as predicted path. Algorithm checks if $\mathcal{F}$=0, it returns no, so it makes a recursive call by the new created query object which equal to $path_2\{e_{11}, e_9, e_7, e_{16}, e_{19}\}$, and the $\mathcal{F}$=1. Then, algorithm identifies the new compared trajectory, assume is $Compared2_\tau(e_3, e_1, e_2)$. Next, the algorithm gets the connected edges to the current node of the query object which are $\{e_{20}, e_{18}\}$. Accordingly, the algorithm creates 2 paths: $path_1\{e_{11}, e_9, e_7, e_{16}, e_{19}, e_{20}\}$,$path_2\{e_{11}, e_9, e_7, e_{16}, e_{19}, e_{18}\}$. Then, the algorithm computes the similarity between each path and $Compared2_\tau$, as a result, the algorithm returns $path_1\{e_{11}, e_9, e_7, e_{16}, e_{19}, e_{20}\}$ as predicted path. The algorithm checks if $\mathcal{F}$=0, it returns yes, so it exit and returns $path_1\{e_{11}, e_9, e_7, e_{16}, e_{19}, e_{20}\}$ as expected path

2) Known Destinations: This section presents the algorithm and illustrative example for case 2; vagrant objects know their destinations and need to reach them.

Algorithm. Algorithm 3 illustrates the pseudo code of the proposed Handler Module that handle case 2 defined before for vagrant objects. The algorithm takes four input parameters, (a) Query object's trajectory Q_τ, (b) Query object's destination $Dest$, (c) Moving object's trajectories $Trajectories_{Current}$, and (d)Distance d. In line 12, in each iteration, the algorithm gets the compared trajectory from the objects currently moved with the query object by using the R-tree within specific distance. In line 13, the algorithm identifies the current edge of the query object, obtains the edges directly connected to the current node and saved them in $edgesList$. Then, the algorithm iterates over the $edgesList$, and for each iterated edge, the algorithm append this edge to the query object trajectory, then consider this as new path. Each path is saved to $PathList$ and algorithm considers it as fetched path. After that, in line 19, the algorithm iterates over the $PathList$, and computes the similarity between each path and the compared trajectory. Hence, the algorithm makes the most similar path as the predicted path. Besides this, the algorithm considers it as a new query object's trajectory, (lines 19-23). Finally, the algorithm checks if the current edge of the query object is same as the destination required, if yes the algorithm returns the predicted path; otherwise, the algorithm makes a recursive calls until reach to its destination (Lines 24-28).

Example. Assume that the query object's trajectory is $\tau_7(e_{11}, e_9, e_7, e_{16})$, assume the destination needs to be predicted e_{20}, assume that $SimilarTrajectoriesList$ S is ϕ, and assume d is 1 Kilo Meter. First, the algorithm gets the compared trajectory $Compared1_\tau$, say for example (e_3, e_1, e_2). Second, gets the connected edges to the current node of the query object which are e_{14}, e_{19}, e_{17}. Accordingly, the algorithm creates 3 paths: $path_1\{e_{11}, e_9, e_7, e_{16}, e_{14}\}$,$path_2\{e_{11}, e_9, e_7, e_{16}, e_{19}\}$, and $path_3\{e_{11}, e_9, e_7, e_{16}, e_{17}\}$. After that, the algorithm computes the similarity between each path and $Compared1_\tau$, as result algorithm returns $path_2$ as predicted path. Algorithm checks if the query object's current edge equal to e_{20}, it returns no, so it make a recursive call by the new created query object which equal to $path_2\{e_{11}, e_9, e_7, e_{16}, e_{19}\}$. Then, algorithm identifies the new compared trajectory, assume is $Compared2_\tau(e_3, e_1, e_2)$. Next, the algorithm gets the connected edges to the current node of the query object which are $\{e_{20}, e_{18}\}$. Accordingly, the algorithm creates 2 paths: $path_1\{e_{11}, e_9, e_7, e_{16}, e_{19}, e_{20}\}$,$path_2\{e_{11}, e_9, e_7, e_{16}, e_{19}, e_{18}\}$. Then, the algorithm computes the similarity between each path and $Compared2_\tau$, as a result, the algorithm returns $path_1\{e_{11}, e_9, e_7, e_{16}, e_{19}, e_{20}\}$ as predicted path. The algorithm checks if if the query object's current edge equal to e_{20}, it returns yes, so it exit and returns $path_1\{e_{11}, e_9, e_7, e_{16}, e_{19}, e_{20}\}$ as expected path.

Algorithm 3 known Destinations Handler

1: **procedure** GET_FUTURE_PATH
2: **INPUT:** Query object's trajectory Q_τ, Query object's destination $Dest$, Moving object's trajectories $Trajectories_{Current}$, Distance d
3: **OUTPUT:** Return $Predicted\_Path$
4: /* Set compared Trajectory */
5: **SET** $Compared_\tau \leftarrow \phi$
6: /* Last edge in query object trajectory */
7: **SET** $Q(CurrentEdge)_\tau \leftarrow \phi$
8: /* edges connected direct to query object current node */
9: **SET** $edgesList \leftarrow \phi$
10: /* All expected paths */
11: **SET** $PathList \leftarrow \phi$
12: $Compared_\tau$ = call R-Tree to get nearest Trajectory to Q_τ from $Trajectories_{Current}$ within d
13: $Q(CurrentEdge)_\tau$ = Current edge of Q_τ
14: *Put in edgesList edges connected by* $Q(CurrentEdge)_\tau$
15: **for** ***each*** *edge* e_i *in* $edgesList$ **do**
16: *FetchedPath=* Q_τ + e_i
17: *Add* $FetchedPath$ *to* $PathList$
18: **end for**
19: **for** ***each*** $FetchedPath$ *in* $PathList$ **do**
20: *ComputeSimilarity(*$FetchedPath$*,*$Compared_\tau$*)*
21: *Set Most Similar FetchedPath as* Q_τ
22: *Set Most Similar FetchedPath as* $Predicted\_Path$
23: **end for**
24: **if** $Q(CurrentEdge)_\tau = Dest$ **then**
25: *Exit and return* $Predicted\_Path$
26: **else**
27: *Get_Future_Path(*$Q_\tau, Dest, Trajectories_{Current}$*)*
28: **end if**
29: **end procedure**

V. EXPERIMENTS

This section evaluates experimentally the proposed $HarmonyMoves$ system.

A. Experimental Setup

Data-sets. All the experiments introduced in this work use real trajectories data-set collected by Microsoft Research team in the Geo-life project [13] covering the period between

April 2007 to August 2012. Furthermore, these trajectories are splitted into small trajectories with an average length of each one 10 road segments. Thus, the total number of trajectories is 5000. All trajectories GPS data points (longitude, latitude) are map-matched to road segments along the road network. The road network data is captured from OpenStreetMap [18]. Road network data in this work represents Hamilton city in the USA. The map-matching algorithm is out the scope of this paper [12].
Experimental Settings. The prediction function employed inside $HarmonyMoves$,R-Tree module, and hausdorff distance functions are implemented using Java with JDK 1.9 inside eclipse PHOTONID IDE. All experiments are accomplished on a PC with Intel(R) core(tm) i7-4770 cpu @3.40ghzM, and running on the linux ubuntu operating system.

B. HarmonyMoves Evaluation and Results

We next report our findings

Exp 1:- Comparison between HarmonyMoves and other baseline models. In this set of experiments, Fig. 4 compares 2 baseline prediction techniques of by $HarmonyMoves$. This is set of experiments choose RandomGuess and RMF for comparison. RMF is an HMM model-based path prediction method , which computes motion function to capture movements. This experiment proves that the proposed system $HarmonyMoves$ outperform other methods in prediction accuracy. The justification behind this is the $HarmonyMoves$ depends on its work on novel similarity function the prune irregularities in input data before the training process.

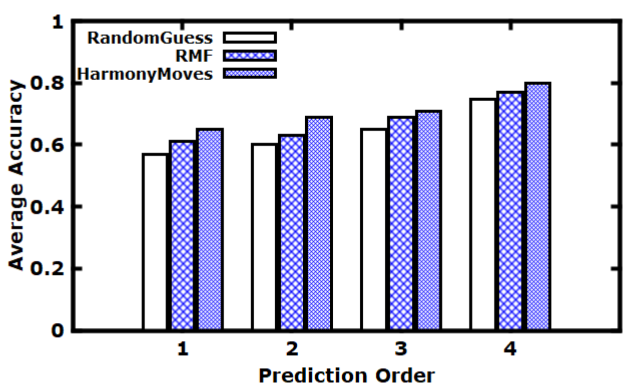


Fig. 4. Comparison between baseline methods

Exp 2:- Accuracy evaluation of vagrant objects VS. normal objects.This set of experiments investigates the impact of increasing number of moving objects on both vagrant objects and normal objects. Fig. 5 shows that the normal objects always achieve higher prediction accuracy more than vagrant objects. The justification behind this is that the normal objects have more objects moving similar like them, and the prediction done through high similarity value, as result prediction in efficient and easily manner.

Exp 3:- Impact of increasing similar objects on normal query objects. This set of experiments studies the impact of increasing number of similar moving objects on the normal

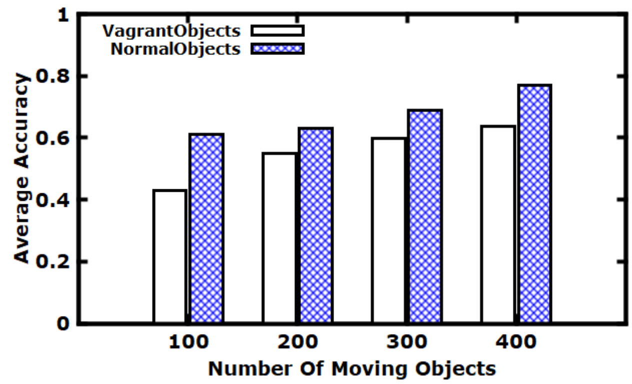


Fig. 5. Vagrant Objects VS. Normal Objects

query objects. It is observed from Fig. 6 that the increase in similar moving objects number, the total increase on the average prediction accuracy. The justification behind this is that the increased number of similar objects will make the prediction HMM model trains on more input data similar to each other and make the model perform efficiently.

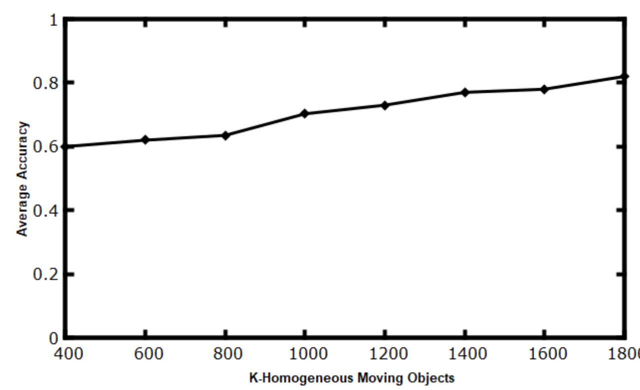


Fig. 6. Impact of increasing similar objects on Normal query objects

Exp 4:- Impact of increasing number of future steps on normal and vagrant objects. In this set of experiments, Fig. 5 compares the normal objects with the vagrant objects according to the CPU processing time when they are need to predict the future steps. It is observed that normal objects consumes less CPU time than vagrant objects during the prediction process. The justification behind this is that the normal objects move similar like other objects and the hash-map created during the prediction can contain cashed values before which can be accessed by O(1). Additionally, vagrant objects make many recursive calls and need each time to identify compared trip to predict next road or final destination.

VI. Conclusion

In this paper, we presented the *HarmonyMoves* system that predicts the future trajectory of a moving query object when the object's movement history or past trajectories is absent. In addition, *HarmonyMoves* helps moving objects that lost their ways to go their destinations or follow recognized routes. The

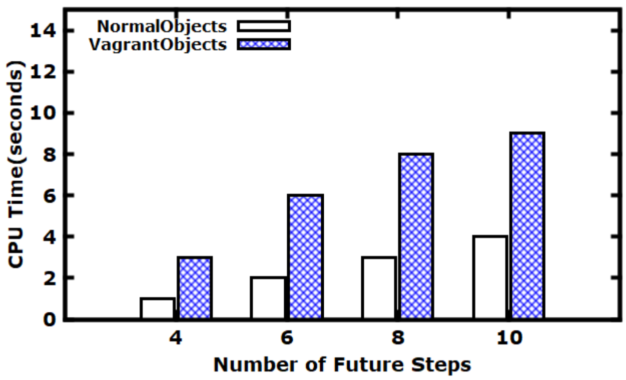


Fig. 7. Impact of increasing number of future steps

system leverages the harmony between the query object and other moving objects that are existing in the space at the same time, explores similarities in their trajectories. If the harmony exist, *HarmonyMoves* utilizes these similarities to predict the query object's future paths. Otherwise, *HarmonyMoves* try to get the nearest object within specific distance to the query object and the then utilize the nearest object's trajectory to guide object that lost their ways. The system composed of three main components; Harmony Checker Module, Normal Objects Handler Module, and Vagrant Objects Handler Module. Experiments showed that *HarmonyMoves* provided accurate prediction results and achieved high performance.

In the future work, we plan to make *HarmonyMoves* able to handle huge number of end-users who submit several predictive queries on the cloud-based through big data frameworks, like Apache-Spark. This makes *HarmonyMoves* more generic and scalable framework.

REFERENCES

[1] Abdeltawab M. Hendawi, Jie Bao, and Mohamed F. Mokbel. iRoad: A Framework For Scalable Predictive Query Processing On Road Networks. *In Proceedings of the International Conference on Very Large Databases, VLDB 2013, Riva Del Garda, Italy*, August, 2013.

[2] Francisco Neto, Cláudio Baptista, and Claudio Campelo. Prediction of Destinations and Routes in Urban Trips with Automated Identification of Place Types and Stay Points. *Brazilian Journal of Cartography*, 2016.

[3] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968), 2010.

[4] Rui Zhang, H.Jagadish, Bing Dai, and Kotagiri Ramamohanarao. Optimized algorithms for predictive range and KNN queries on moving objects. *Information Systems*, 35(8), 2010.

[5] John Krumm,Robert Gruen,and Daniel Delling. From destination prediction to route prediction. *Journal of Location Based Services*, 2013.

[6] Jon Froehlich and John Krumm. Route Prediction from Trip Observations. *SAE Technical Paper Series*, 2008.

[7] Abdeltawab Hendawi and Mohamed Mokbel. Panda. *Proceedings of the 20th International Conference on Advances in Geographic Information Systems-SIGSPATIAL 12*, 2012.

[8] Abdeltawab Hendawi and Mohamed Mokbel. Predictive Spatio-Temporal Queries: A Comprehensive Survey and Future Directions. *In Proceeding of the first ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, MobiGIS 2012, Redondo Beach, CA,*, November 2012.

[9] Heng Shen, Hoyoung Jeung, Qing Liu and Xiaofang Zhou. A Hybrid Prediction Model for Moving Objects. *2008 IEEE 24th International Conference on Data Engineering*, 2008.

[10] Hassan Karimi and Xiong Liu. A predictive location model for location-based services. *In Proceedings of the ACM-GIS'03*, 2003.

[11] John Krumm. A Markov Model for Driver Turn Prediction. *SAE Technical Paper Series*, 2008.

[12] John Krumm, Julie Letchner, and Eric Horvitz. Map Matching with Travel Time Constraints. *SAE Technical Paper Series*, 2007.

[13] Yu Zheng, Xing Xie, Wei-Ying Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[14] Irmtraud Meyer. Hidden Markov Model (HMM, Hidden Semi-Markov Models, Profile Hidden Markov Models, Training of Hidden Markov Models, Dynamic Programming, Pair Hidden Markov Models). *Dictionary of Bioinformatics and Computational Biology*, 2004.

[15] Klaus U.Schulz, Stoyan Mihov. Fast string correction with Levenshtein automata. *International Journal on Document Analysis and Recognition*, 2002.

[16] Abdeltawab Hendawi, Jie Bao, Mohamed Mokbel, and Mohamed Ali. Predictive Tree: An Efficient Index for Predictive Queries on Road Networks. *2015 IEEE 31st International Conference on Data Engineering*, 2015.

[17] Emilian Necula. Dynamic traffic flow prediction based on GPS Data. *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, 2014.

[18] OpenStreetMap. https://www.openstreetmap.org/, accessed on april 2019.

[19] Yufei Tao, Dimitris Papadias and Jimeng Sun. The TPR*-tree: an optimized spatio-temporal access method for predictive queries. *Proceeding VLDB '03 Proceedings of the 29th international conference on Very large data bases*, 2003.

[20] Katerina Raptopoulou, Apostolos Papadopoulos, and Yannis Manolopoulos. Fast Nearest-Neighbor Query Processing in Moving-Object Databases. *GeoInformatica*, 7(2):113–137, 2003.

[21] Philippe C.Besse,Brendan Guillouet,Jean-Michel Loubes,and Francois Royer. Destination Prediction by Trajectory Distribution Based Model. *IEEE Transactions on Intelligent Transportation Systems*, pages 2470–2481, 2018.

[22] Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider and Bernhar Seeger. The R*-tree: an efficient and robust access method for points and rectangles. *ACM SIGMOD Record*, 19(2):220–231, 1990.

[23] B. Sendov. Hausdorff Distance. *Hausdorff Approximations*, pages 23–48, 1990.

[24] Hoda Mokhtar, Jianwen Su, and Oscar Ibarra. Map Matching with Travel Time Constraints. *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS 02*, 2002.

[25] Xiaofeng Xu, Li Xiong, Vaidy Sunderam, and Yonghui Xiao. A Markov Chain Based Pruning Method for Predictive Range Queries. *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems-GIS 16*, 2016.

[26] J. Sun, Dimitris Papadias, Yufei Tao, and Bin Liu. Querying about the past, the present, and the future in spatio-temporal databases. *Proceedings. 20th International Conference on Data Engineering*, 2004.

[27] Ning Ye,Zhong-Qin Wang,Reza Malekian,Qiaomin Lin,and Ru-Chuan Wang. A Method for Driving Route Predictions Based on Hidden Markov Model. *Mathematical Problems in Engineering*, 2015.

[28] Yisong Zheng, Qian Wang, Wangling, and Mo Kuang. Research into the driver's route choice under existing real-time traffic information. *2008 IEEE International Conference on Industrial Engineering and Engineering Management*, 2008.

[29] R. Simmons, B. Browning, Yilu Zhang, and V. Sadekar. Learning to Predict Driver Route and Destination Intent. *2006 IEEE Intelligent Transportation Systems Conference*, 2006.

[30] Shaojie Qiao, Nan Han, William Zhu, and Louis Gutierrez. TraPlan: an effective three-in-one trajectory-prediction model in transportation networks. *IEEE Transactions on Intelligent Transportation Systems*, 16(3), 2015.

A Survey on Cloud Data Security using Image Steganography

Afrah Albalawi[1], Nermin Hamza[2]
Faculty of Computing and Information Technology, King Abdulaziz University
Jeddah, Saudi Arabia[1,2]
Faculty of Graduate Studies for Statistical Research, Cairo University[2]

***Abstract*—Now-a-days, cloud computing proved its importance where it is being used by small and big organizations. The importance of cloud computing is due to the various services provided by the cloud. One of these services is storage as a service (SaaS) which allows users to store their data in the cloud databases. The drawback of this service is the security challenge since a third party manages the data. The users need to feel safe to store their data in the cloud. Consequently, we need for models that will enhance the data security. The image steganography is a way to protect data from unauthorized access. Image steganography allows users to conceal secret data in a cover image. In this paper, we review and compare some of the recent works proposed to protect cloud data using image steganography. The first comparison of models based on the algorithms they used, advantages and drawbacks. The second comparison of the models based on the aims of steganography: quality where the model produces a stego-image with high quality, security where the secret data is difficult to detect and capacity where the model allows to hide large amounts of data.**

Keywords*—*Security; cloud computing; image steganography; data hiding; data storage

I. Introduction

Cloud computing provides flexible services for users by combining many of resources and applications based on a pay-as-you-need concept [1]. One of the services provided by the cloud is store data in the cloud. This service provides fast distribution, low-cost and reliability [2].

When storing data in cloud storage, storage devices has vulnerability to internal leakage, hacking and other reasons that may lead to lose data confidentiality [3].

Some of data stored in the cloud are very sensitive data, such as banking and government information, which must be protected against unauthorized people including the cloud service provider [2].

There are many researches that use cryptography techniques to protect the cloud confidentiality of the data [1], but the main disadvantage of encryption is although the data is encrypted and became unreadable, it is still exists as a secret data. The attacker could decrypt the data if he has enough time [4]. Steganography is a way to solve this problem since it allows the user to hide data into other object such as text, image, audio or video, these techniques will increase the sensitive data security [1]. In this paper, we focus on the image steganography to protect cloud data. Fig. 1 illustrate the usage of image steganography in cloud environment.

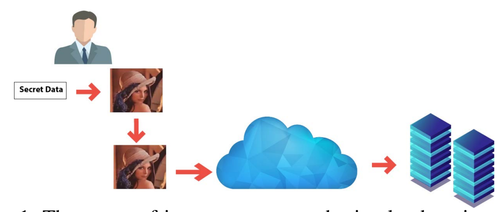


Fig. 1. The usage of image steganography in cloud environment

In this paper, we check the existing cloud data security techniques used image steganography. This paper is structured as follows. Section II, present an overview of cloud computing. Section III, give an overview of steganography. In Section IV, we introduce an overview of image steganography. In Section V, we review some of recent techniques of cloud data security using image steganography. In Section VI, we compare the techniques based on different aspects and discuss the current status. In Section VII, we discuss the future works.

II. Cloud Computing Overview

In this section, we give an overview of the cloud computing, service models, deployment models and security requirements of cloud computing.

Cloud computing provides IT services to users over the Internet. The NIST defined cloud computing as “a model for enabling ubiquitous, convenient, on-demand network access to shared pool of configurable computing resources (e.g., networks, servers, storage applications and services) that can be rapidly provisioned and released with minimal management effort and or service interaction” [5].

A. Service Model of Cloud Computing

- Software as a service (SaaS): User can only use the applications provided by the provider without ability to manage the applications [6].
- Platform as a service (PaaS): User creates applications on the cloud infrastructure and the user will be able to deploy and manage the applications [6].
- Infrastructure as a service (IaaS): User will provide the fundamental computing resources, such as networks, storage and processing [6].
- File storage as a service (FSaaS): The cloud provides the ability to store, manage and access the data from

an interface of browser. the cloud provider holds the maintenance responsibility and oversees the infrastructure storage [7].

B. Deployment Model of Cloud Computing

- Private cloud: Cloud service provider makes the resources and applications available to cloud users. The users must subscribe to get the benefits of the resources, and they will pay based on the subscription [6].
- Public cloud: Users use the resources dynamically over the Internet, and they will pay based on their use [6].
- Hybrid cloud: It consists of distributed private clouds linked together and have a central management. The payment system in this model is complex [6].

C. Cloud Computing Security Requirements

- Audit: It includes authentications and authorisation, to ensure user's identity by implementing a strong verification process [8].
- Confidentiality: Protect data stored in the database from unauthorized users [8].
- Integrity: It is used to ensure the data consistency, and to protect data from iteration [9].

III. Steganography Overview

In this section, we presents an overview of the steganography, its types, its objectives.

Steganography is the science of hiding the secret data in a multimedia file [10]. Steganography as a word is combine of two Greek words "Sregano" and "Graphy", and the meaning is "cover writing" [11].

A. Types of Steganography

- Text steganography: Use text file to hide secret data [12].
- Image steganography: Hide secret data in a cover image [13].
- Audio steganography: Use an audio file to conceal secret data [13].
- Video steganography: Hide secret data in a video file [14].
- DNA-based steganography: Employ randomness of DNA to embed secret data [15].
- Protocol steganography: Hide secret data in network protocol such as, IP, TCP and UDP [16].

B. Objectives of Steganography

- Security: The attacker unable to detect the secret data [7].
- Payload (Capacity): Allow to hide large amount of data into the cover object [7].
- Invisibility (Quality): The changes in the cover object undetectable by the Human Visual System (HVS) [7].

IV. Image Steganography Overview

This section, provides an overview of image steganography, some techniques of image steganography and types of images.

The image steganography is the process of hiding the secret data in a cover image to produce a stego image [7].

A. Some of Image Steganography Techniques:

- Least Significant Bit (LSB) based Steganography: Hide the bits of secret data in the LSB of the cover image. This technique is the most popular used [17].
- Discrete Cosine Transform (DCT): Use subdivision of quantized DCT coefficient to hide the secret data [17].
- Discrete Wavelet Transform (DWT): It is used to decompress the image mathematically into a set of wavelet [7]. This technique used for medical and military applications [17].

B. Types of Images

- The binary images: consists of black and white pixels [7].
- The grayscale images: consists of pixels with shades of gray colors [7].
- The color images: uses some integration of red, green and blue to specify the pixels' colors [7].

V. Current Works Proposed for Cloud Data Security using Image Steganography

In this section, we review some works proposed for cloud data security using image steganography.

Mohis and Devipriya in [2], proposed an improved approach that increases the security of public cloud data by using mediated certificateless public key encryption (MCL-PKE) and LSB steganography algorithm. The proposed system consists of three modules: registration module, cloud module and embedding module. In the registration module, the user registers to the cloud and generates public and private keys, keep the private key for the users and transfer the public key to the Key Generation Centre (KGC). In the cloud module, if the user requests the data the Security Mediator (SEM) check if the user legitimate it will decrypt partially the data and will provide it to the user, then the user fully decrypts it using the private key. In the embedding module, the user before storing the data in the cloud he will embed the sensitive data within an image. The authors compared the proposed approach with other system. The proposed approach reduces overhead at the owner side, and reduces unauthorized access on the data. This technique does not produce high quality stego image and does not allow to hide large amount of data.

Ebrahim et al. [1] combined encryption and steganography to prevent unauthorized access to cloud data. In the proposed model, there are three phases. The first phase, compute hash value of secret data using SHA-256, then use RSA to encrypt the hash value and session key. The second phase, use AES-256 to encrypt the secret data. The third phase, use advanced LSB algorithm to hide encrypted data in a cover image. The authors were evaluated the proposed model and compared it with other models. The result shows this model provides security against cryptanalysis and steganalysis attacks and stetisstical changes, and produces a stego image with high quality.

Seshubhavan et al. in [18], used steganography and genetic algorithms to secure the data in the cloud. The proposed technique tries to insert the secret data in suitable pixels in the cover image without affecting the characteristics of the cover image. This technique work only on the grayscale image. Therefore, if the cover image is colored image convert it to gray scale image, then extract the least significant bit and most significant bit and convert them to 0's and 1's array. Use the AES algorithm to encrypt the secret data and the key converted to 0's and 1's array. The two arrays combined and split into R Block, and L Block. These segments are applying to genetic algorithm to produce an address block, which is used to embed the secret data in the cover image and produce the stego image that will store in the cloud database. This algorithm compared with other existing algorithms. The result shows that, the proposed algorithm is better quality, but does not provide high capacity payload.

Rahman et al. [19] proposed a new combination of encryption and steganography to secure cloud data. They used blowfish algorithm to encrypt secret data, to embed encrypted data in a cover image E-LSB algorithm is used, and to preserve the integrity of produced stego image they used SHA-256. The analysis of the proposed model presents the model provides security against statistical and visual attacks.

Suneetha and Kumar in [20] have improved the security of cloud data by using partition random edge-based technique for image steganography. They supposed this technique will help to reduce changes between cover image and produced a stego image. In the embedding process, convert the cover image into grayscale image and portion it into 9 partitions. Then, use Canny edge detection method to identify the edge pixels and select the prime number of random pixels of an image. After that encrypt the secret data and embed the key in the selected pixels. The authors compared their method with others existing methods and the result shows that, this method is better and works on different types of data. It provides security against steganalysis attack. This work focuses on security and quality, but ignores the amount of data can be embedded in the cover image.

Kumar and Suneetha [21] used image segmentation along with image steganography to increase the security of data in cloud environment. To embed secret data in a cover image covert the cover image into black and white or grayscale image, then apply the image segmentation technique to identify and extract the iris part of the cover image. After that use Canny edge detection to select edge pixels of inner and outer circle, and use RSA algorithm to encrypt the secret data. Hide the secret key in the selected pixels and store the stego image in the cloud. The authors were analyzed the technique and the result shows that this technique provides better security than others existing techniques based on steganography and segmentation.

Shanthakumari and Malliga [22] proposed a combination of International Data Encryption Standard Algorithm (IDEA) and Least Significant Bit Grouping (LSBG) algorithm to improve security and capacity of data embedding to the cover image. In the embedding phase, the IDEA algorithm performed to encrypt secret data, then LSBG is applying to embed the encrypted data into cover image and produce stego image which is uploaded to the cloud. In the extracting phase, download the stego image from the cloud and use LSBG to extract the secret data, then perform IDEA decryption to decrypt the secret data. The authors were evaluated the proposed technique and compared it with other techniques. The result shows this technique provides good security for secret data and produces stego image with high quality and increase the embedding capacity.

VI. DISCUSSION

In this section, we compare the current techniques based on different aspects and discuss the current status.

Table I shows a comparison of the reviewed techniques based on the algorithms they used, advantages and drawbacks. From Table I, we conclude there is no technique totally strong without weaknesses, each technique has its own strengths and weaknesses.

In Table II, we compare the current techniques based on steganography objectives: security, capacity and quality.

From Table II, we can conclude all proposed techniques satisfies the security objective, and five of them produces a stego image with high quality, but only one technique allows to hide large amounts of data.

The reviewed techniques works on different types of images where [18] and [20] suitable for a grayscale image, [21] works on black and white or grayscale images and some suitable for color image such as [1], [19] and [22].

VII. FUTURE WORK

for the future work, we intend to improve efficient solutions that satisfy the objectives of steganography. These solutions should allow to hide large amounts of data, produce a stego image with high quality and the hidden data must be undetectable.

VIII. CONCLUSION

Cloud computing provides many benefits to the users but it has security challenges. Image steganography is a way to protect secret data in the cloud by hide the secret data in a cover image. This paper, present a review of some

TABLE I. COMPARING CLOUD DATA SECURITY ALGORITHM USED IMAGE STEGANOGRAPHY

Author	Algorithms used	Advantages	Drawbacks
[2]	-Mediated certificateledd public key encryption (MCL-PKE). -LSB steganography.	-Reduce overhead at the owner side. -Protect the secret data against unautherized access.	-Does not produce a stego image with high quality. -Does not allow to hide large amount of data. -Expensive and difficult. -Does not provide how to instant revocation when desired.
[1]	- SHA-256 - RSA encryption algorithm - AES-256 encryption algorithm advanced LSB steganography algorithm.	- Allows to embed any type of data in any format of image. - Produces stego image with high quality. - Protects cloud data against attacks and statistical changes.	- Capacity of embedded data depends on the size of cover image. - Time consuming.
[18]	-Genetic algorithm. -AES encryption algorithm.	Allows to insert secret data in cover image without affecting the characteristic of the cover image.	-use only with gray scale images.
[19]	- Blowfish encryption algorithm - E-LSB steganography algorithm - SHA-256	Provides security against statistical and visual attacks.	Ignores the quality of stego image and capacity of embedded data.
[20]	-Partition random edge-based. -Canny edge detection. -Encryption algorithm.	-Reduces changes between cover image and stego image produced. -Suitable for different type of data. -Provide security against steganalysis attach.	-Used only for grayscale images. -Time consuming. -High requirements on memory.
[21]	-Image segmentation. -Canny edge detection. -RSA encryption algorithm.	increase the security.	-Using for black and white or grayscale images. -Time consuming. -High requirements on memory.
[22]	-IDEA encryption technique. -LSBG steganography algorithm.	- Increase the security. - Produce stego image with high quality - Allow to hide large amount of data.	IDEA uses large number of weak keys.

TABLE II. COMPARING THE TECHNIQUES BASED ON STEGANOGRAPHY OBJECTIVES

Author	*Security*	*Quality*	*Capacity*
[2]	Yes	No	No
[1]	Yes	Yes	No
[18]	Yes	Yes	No
[19]	Yes	No	No
[20]	Yes	Yes	No
[21]	Yes	Yes	No
[22]	Yes	Yes	Yes

recently proposed techniques for cloud data security using image steganography. We compared these techniques based on algorithms they used, advantages and drawbacks, and based on the objectives of steganography. We concluded each technique has its own advantages and weaknesses which make it difficult to choose one technique as the best solution.

REFERENCES

[1] M. A. Ebrahim, I. A. El-Maddah, and H. K. Mohamed, “Hybrid model for cloud data security using steganography,” in *2017 12th International Conference on Computer Engineering and Systems (ICCES)*. IEEE, 2017, pp. 135–140.

[2] M. Mohis and V. Devipriya, “An improved approach for enhancing public cloud data security through steganographic technique,” in *2016 International Conference on Inventive Computation Technologies (ICICT)*, vol. 3. IEEE, 2016, pp. 1–5.

[3] R. Wang, “Research on data security technology based on cloud storage,” *Procedia engineering*, vol. 174, pp. 1340–1355, 2017.

[4] B. Datta and S. K. Bandyopadhyay, “Cloud steganography-a review,” *Journal for Research— Volume*, vol. 2, no. 01, 2016.

[5] P. R. Kumar, P. H. Raj, and P. Jelciana, “Exploring data security issues and solutions in cloud computing,” *Procedia Computer Science*, vol. 125, pp. 691–697, 2018.

[6] G. Ramachandra, M. Iftikhar, and F. A. Khan, “A comprehensive survey on security in cloud computing,” *Procedia Computer Science*, vol. 110, pp. 465–472, 2017.

[7] A. Y. AlKhamese, W. R. Shabana, and I. M. Hanafy, “Data security in cloud computing using steganography: A review,” in *2019 International Conference on Innovative Trends in Computer Engineering (ITCE)*. IEEE, 2019, pp. 549–558.

[8] M. Yesilyurt and Y. Yalman, “New approach for ensuring cloud computing security: using data hiding methods,” *Sādhanā*, vol. 41, no. 11, pp. 1289–1298, 2016.

[9] Y. AlHumaidan, L. AlAjmi, M. Aljamea, and M. Mahmud, “Analysis of cloud computing security in perspective of saudi arabia,” in *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2018, pp. 1–4.

[10] A. S. Ansari, M. S. Mohammadi, and M. T. Parvez, “A comparative study of recent steganography techniques for multiple image formats,” *International Journal of Computer Network and Information Security*, vol. 11, no. 1, p. 11, 2019.

[11] I. J. Kadhim, P. Premaratne, P. J. Vial, and B. Halloran, “Comprehensive survey of image steganography: Techniques, evaluations, and trends in future research,” *Neurocomputing*, vol. 335, pp. 299–326, 2019.

[12] A. Beroual and I. F. Al-Shaikhli, “A review of steganographic methods and techniques,” *International Journal on Perceptive and Cognitive Computing*, vol. 4, no. 1, pp. 1–6, 2018.

[13] A. Arya and S. Soni, “A literature review on various recent steganography techniques,” *International Journal on Future Revolution in Computer Science& Communication Engineering*, vol. 4, pp. 143–149, 2018.

[14] M. HASHIM, M. RAHIM, M. SHAFRY, and A. A. ALWAN, "A review and open issues of multifarious image steganography techniques in spatial domain." *Journal of Theoretical & Applied Information Technology*, vol. 96, no. 4, 2018.

[15] M. Hussain, A. W. A. Wahab, Y. I. B. Idris, A. T. Ho, and K.-H. Jung, "Image steganography in spatial domain: A survey," *Signal Processing: Image Communication*, vol. 65, pp. 46–66, 2018.

[16] W. A. Awadh, A. S. Hashim, and A. K. Hamoud, "A review of various steganography techniques in cloud computing," *University of Thi-Qar Journal of Science*, vol. 7, no. 1, pp. 113–119, 2019.

[17] S. Jeevitha and N. Amutha Prabha, "A comprehensive review on steganographic techniques and implementation," *ARPN Journal of Engineering and Applied Sciences*, vol. 13, p. 17, 2018.

[18] M. Seshubhavan, D. Suneetha, and S. A. Varma, "A novel approach for data security for cloud data using image steganography and genetic algorithms." *Grenze International Journal of Engineering & Technology (GIJET)*, vol. 4, no. 3, p. 137, 2018. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=134178986&lang=ar&site=eds-live

[19] M. O. Rahman, M. K. Hossen, M. G. Morsad, and A. Chandra, "An approach for enhancing security of cloud data using cryptography and steganography with e-lsb encoding," *IJCSNS*, vol. 18, no. 9, p. 85, 2018.

[20] D. Suneetha and R. K. Kumar, "Enhancement of security for cloud data using partition-based steganography," in *Proceedings of the 2nd International Conference on Data Engineering and Communication Technology*. Springer, 2019, pp. 201–209.

[21] R. K. Kumar and D. Suneetha, "A novel approach for data security in cloud environment using image segmentation and image steganography," in *Information Systems Design and Intelligent Applications*. Springer, 2019, pp. 75–82.

[22] R. Shanthakumari and S. Malliga, "Dual-layer security of image steganography based on idea and lsbg algorithm in the cloud environment," *Sādhanā*, vol. 44, no. 5, p. 119, 2019.

Developing Decision Tree based Models in Combination with Filter Feature Selection Methods for Direct Marketing

Ruba Obiedat
King Abdullah II School of Information Technology
The University of Jordan
Amman, Jordan

***Abstract*—Direct Marketing is a form of advertising strategies which aims to communicate directly with the most potential customers for a certain product using the most appropriate communication channel. Banks are spending a huge amount of money on their marketing campaigns, so they are increasingly interested in this topic in order to maximize the efficiency of their campaigns, especially with the existence of high competition in the market. All marketing campaigns are highly dependent on the huge amount of available data about customers. Thus special Data Mining techniques are needed in order to analyze these data, predict campaigns efficiency and give decision makers indications regarding the main marketing features affecting the marketing success. This paper focuses on four popular and common Decision Tree (DT) algorithms: SimpleCart, C4.5, RepTree and Random Tree. DT is chosen because the generated models are in the form of IF-THEN rules which are easy to understand by decision makers with poor technical background in banks and other financial institutions. Data was taken from a Portuguese bank direct marketing campaign. A filter-based Feature selection is applied in the study to improve the performance of the classification. Results show that SimpleCart has the best results in predicting the campaigns success. Another interesting finding that the five most significant features influencing the direct marketing campaign success to be focused on by decision makers are: Call duration, offered interest rate, number of employees making the contacts, customer confidence and changes in the prices levels.**

Keywords*—Direct marketing; data mining; decision tree; simpleCart; C4.5; reptree; random tree; weka; confusion matrix; class-imbalance

I. INTRODUCTION

Direct marketing has become a trend topic for academics and researchers over the past few years due to high competition between companies, increasing marketing campaigns costs and the changing demands of customers which make it hard to predict [29] [22]. Direct marketing is about finding the most potential customers for a certain product based on their characteristics, interests, behavior and needs, then trying to make customized marketing campaigns for these customers. All industries aim to increase their returns of marketing campaigns and their sales consequently through using the right marketing channels and techniques directed to the right customers at the right time [15]. Banks present one of the major sectors which have a great pressure to increase profits and reduce costs through using the right marketing strategies [17].

There are two approaches for promotions: mass marketing and direct marketing. Mass marketing uses the traditional media for promotion such as television, radio, newspapers and broadcast messages to be distributed randomly without any customization [15], [12]. This type of marketing becomes less effective with time because of the great competition and the large number of available products these days along with its high cost. Usually the response rate which presents the percentage of customers who are influenced by the marketing and actually buy the promoted products does not exceed 1% which is considered a very low percentage. It is to be noted that, industries hope to increase this rate using direct marketing [13][29] [22].

Data mining techniques, machine learning and business intelligence present important models which can be used for direct marketing since there is a huge amount of available data about customers stored in the databases [4], [13], [29][32], which makes it impossible to analyze this data manually [15] [2], [20]. This data can be studied and analyzed in order to discover the customers' behavior, interests and pattern of buying. This information presents an important source of data for decision makers to help them predict the most potential customers to focus on with direct marketing and increase the respond rate consequently [13], [29], [2], [12]. This, ultimately leads to better management of the available resources to target these customers [19]. Direct marketing is used widely by many industries especially retailers, banks and insurance companies to promote their product and services such as loans and retirement insurance [13]. The reason why they use it is the massive amount of available data about their customers which is generated on a regular basis in an electronic format [2]. Most of the time, classification data mining approach is applied for this purpose in direct marketing to predict whether the customers are classified as buyers or non-buyers [19]. Nevertheless, the marketers' poor skills and knowledge of the data mining models makes it difficult for them to use these models[29].

This study aims to use a simple and comprehensive data mining model which is easy to be understood by users with little or no technical background, especially that decision makers in this case are usually sales persons and managers who are responsible for the direct marketing decisions and it is hard for them to use, understand and interpret more complex models even if these models have more predictive power. In one way or another, Decision Tree algorithms are the best choice here since the results they give are readable, comprehensive rules which can be translated easily to a natural language as a series of IF-Then statements for marketers instead of black boxes models.

The main problem in using data mining with direct marketing is the high imbalance in the class distribution; as the response rate for these campaigns is less than 1% which presents the positive examples (buyers and respondent) and the rest 99% is identified as negative percentage. Most data mining algorithms do not behave well with this imbalance [13], [19]. Some studies such as [13] proposed using a learning algorithm which not only classifies examples but can also can compute probabilities and rank the example from most likely to least likely buyers. Hence lift analysis was used for evaluation.

This paper is structured as follows: Section II discusses the related works. Section III identifies the methodological approach followed in this research. Experiments and results are discussed in Section IV and finally conclusions are drawn in Section V.

II. Related Work

This section reviews the main studies that discussed the usage of data mining techniques in direct marketing and highlights the main algorithms they applied along with their obtained results.

A two step approach was followed by [13] in order to discuss the data mining methods used for direct marketing. Firstly, Data mining was used to categorize the current customers into likely buyers and non-buyers in order to focus promotion on the likely buyers then apply the chosen data mining algorithms.

Three data sets taken from three different sources were used by the study for direct marketing. Only a small number of customers were identified as buyers. After that they tried to find the potential customers from the current non-buyers. The first data set was taken from a well known Canadian bank using their promotion for loan product. 90.000 records were studied and each customer has 55 attributes and after preprocessing, 62 attributes were used for data mining. The second data set was taken from a major life insurance company using a registered retirement saving campaign. The data set contains 80.000 customers with 7% identified as buyers and each customer has 10 attributes. The third data set belongs to a company that runs a bonus program for 100 sponsors. The data set contains 104.000 customers with 1% responders and each customer has 299 attributes [13].

The study chose Naïve Bayes algorithm and decision tree C4.5 algorithm with a slight modification to produce Certainty Factor (CF). Lift index was used for evaluation. Ada-boost methods of ensembling classifier were applied before applying the learning algorithms. Results shows that data mining can improve the efficiency of direct marketing in terms of the number of respondents and profit [13].

Other studies such as [11] applied data mining technology in the credit card marketing to help banks use the favorable strategy in finding the target clients based on real data taken from Chinese commercial banks. Firstly, they used K-mean clustering to divide the credit card holders, then built four classification models (C5.0, neural network, chi-squared, classification and regression tree). The result revealed that the decision tree is the best model to obtain the necessary features (e.g. monthly income, family size and age) for successful credit card direct marketing.

Furthermore, [4] applied a Multi-Layer Perception Neural Network (MLPNN), Bayesian networks, Logistic Regression (LR) and (C5.0) decision tree in order to increase the efficiency of the marketing campaign. Real-world data of bank deposit was used. Results proved the effectiveness of these algorithm in predicting the best contact channel with the customers for subscribing deposits. Three statistical measures were used for evaluation, which are accuracy, sensitivity and specificity.

The same data set used by this study was collected and used by [17] who applied logistic regression, neural network, decision trees and support vector machine on the data set of the same bank with 22 selected features. Neural Network had the best results regarding the used metrics AUC of 0.8 and LIFT of 0.65. Moreover the results prove that 79% of successful contacts can be achieved by contacting only half of the better classified customers instead of calling all of them. Finally, sensitivity analysis and DT were applied and revealed that three months euribor rate followed by the call direction (inbound or outbound) was the most relevant feature. In addition, [19] also contained real data from a Portuguese bank concerning 17 phone marketing campaigns. Three CRISP-DM iterations were followed. The researchers applied many data mining algorithms such as Naïve Bayes (NB), Decision Trees (DT) and Support Vector Machines (SVM). The results showed that SVM has the highest prediction performance followed by NB and DT respectively. The most relevant feature was the call duration and the month of contact came next. In the same context using the same dataset [18] proposed a divide and conquer strategy using neural network data mining technique in order to divide the problem into a smaller manageable sub-problems. Each sub-problem is characterized by certain features. Experts evaluated the top influential features of the campaign and considered the call direction (inbound/outbound) as the most relevant one.

On the other hand [2] discussed a case study of a rural bank in Ghana. It applied J48 decision tree and Naïve Bayes. The data set contained 1000 instances with 10 features. The experiment found that the DT accuracy was better than NB with 92.5% and 91.6% respectively. Additionally, it identified the number of contacts as the most important attribute for the J48 DT.

Some studies followed two steps analysis starting by clustering the customer according to their characteristics and needs then made the classification models. For example [15] defined a set of users and tried to align them with the most appropriate communication channels and products. It followed two methods which are partitioning and model based prediction. First it clustered the products and channels then used these clusters in order to predict the customers' decision. The best results in term of accuracy and positive ration were obtained using 5 clusters. However, in terms of the classification methods C4.5 decision tree and Naive Bayes were the best. Finally the results showed that the partitioning method alone increased the accuracy, TP and TN values whereas combining the partitioning method with the classification model yielded to higher accuracy.

Other studies followed a comparative approach such as [33], which used a UCI repository data set with 16 attributes and 45,211 instances to compare between different classification techniques in bank direct marketing. The study chose four algorithms which are SVM, LAD-tree algorithm, J48 and Radial Basis Function Network (RBFN). SVM achieved the highest accuracy while RFBN was the worst one with percentages of 86.95% and 74.34%, respectively.

In general it can be noticed that most of the previous works focused on applying different data mining techniques and comparing between them in terms of efficiency. Nevertheless, not much attention has been given to complexity issues which present a serious concern here, since it is difficult for decision makers with little technical background to understand the complex relationships between the considered attributes. Therefore, this work attempts to cover this gap by focusing on applying a simple model which is easy to interpret since the decision makers in this case are managers and sales person who are not technical employees in the first place, which made DT the most appropriate option.

III. Methodology

There are many methodologies that can be adopted for constructing the data mining model. This paper follows a five stage methodology framework that aims to examine and modify the prediction model. This process of data mining is useful, simple and flexible to many people who have fair experience in the field of data mining . Fig. 1 below illustrates the proposed methodology of this research.

The main five ideas this experiment is interested in are shown through the following steps:

- Feature Selection: Also Known as Attribute selection. It is a useful method to reduce the number of attributes by illuminating the irrelevant attributes that do not highly affect the utility of data [10]. Using Feature selection techniques reduces the computation time, simplify the model and reduces the over-fitting. In Weka, there are three options for performing attribute selection which are using the attribute selection tab directly, using a meta-classifier and using the filter approach. This experiment used the meta-classifier option and the select attributes tab to obtain the numerical weight of each attribute.
- Tree based Models Building: This paper discusses four types of decision trees classification algorithms (SimpleCart, C4.5, RepTree and Random Tree). Decision trees are considered one of the most powerful and common tools for classification and prediction. Decision trees produce rules, which can be understood and interpreted easily by humans working in any domain.
- Performance Evaluation: This study has used the most common model evaluation metric such the accuracy, True Positive Rate (Recall), Precision, F-Measure and ROC area which are all derived from the confusion matrix without the need of any manual calculation. Moreover, this study reviews an additional evaluation metrics for the evaluation of a model's efficiency in the presence of highly imbalanced data. It does so by applying the Geometry Mean (G-Mean).
- Feature Analysis: This step compares the achieved result of the feature selection and data reduction techniques of the most important 5,10 and 15 top attributes. This method selects the best 5 attributes from the total 21 attributes.
- Rules Analysis: This presents the last step in the methodology framework in which the most important rules are extracted as a series of IF-THEN statements relying on the tree with the best results. These rules highlight the most significant features to be focused on by decision makers.

A. Constructing the Prediction Model

Decision trees are one of the most commonly used models in machine learning and decision analysis; they help in determining the most successful strategy to reach the target. They are considered a predictive method which can be used for both classification and regression models. Decision trees are a supervised approach that seeks to find the relationship between input attributes and output attribute (class label) for optimal prediction [16].

The idea of the decision trees can be presented as the tree structure, where each node represents an attribute, each branch represents an outcome of the test, and each leaf node denotes a class label. The decision tree classifier traces the path from the root which is the main attribute of the set to the leaf node, which represents the class label [27]. The decision trees algorithm has a statement "if ... then ... else ..." construction which makes it easy to read and interpret. Moreover, Decision trees algorithms have different features which and this difference causes a difference in their results.

In this paper, different decision tree algorithms were used to predict the bank direct marketing campaigns.

which are:

- C4.5: This algorithm is developed by Ross Quinlan and is used to generate a decision tree [5]. It is ad extension of ID3 algorithm that solves most of its problems, like dealing with noise and missing data and it is often used as a statistical classifier. C4.5 builds a decision tree based on the gained information. The attribute with the highest information gain is used as the splitting criteria. Moreover, C4.5 uses Gain Ratio for attribute selection criteria. This method contains two concepts which are Gain and Split Info. In other words, for continuous attributes this selection criteria gives the best result compared to ID3, which is only appropriate for discrete datasets [27]. Nevertheless, C4.5 has few disadvantages like the small variation in data, which causes different decision trees in addition to and the fact that it is not suitable for small training set [5]
- RandomTree: This is a supervised classifier developed by Leo Breiman and Adele Cutler. It can handle both classification and regression problems [8]. During The classification, each input feature is classified with all the trees in the forest,

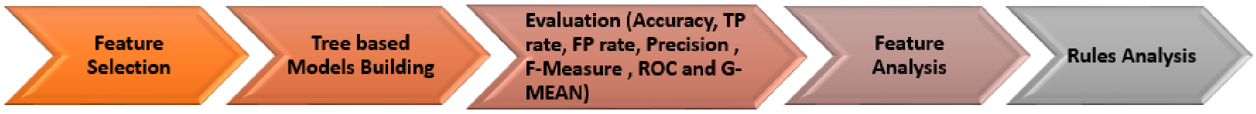


Fig. 1. Proposed Methodology.

and the class label will be the output of the majority. In regression problems, the classifier response is the average of the responses over all the trees in the forest [9].

- SimpleCart: CART is a prediction algorithm that was developed in the early 80s in Southern California by Leo Breiman [8]. It is considered as Classification and Regression Tree that uses historical data in order to generate a binary decision tree. It can operate with categorical or numeric attributes and this distinguishes it from other decision trees methods [26], [8]. One of the advantages of CART method is its strength to outliers. While splitting the algorithm it will isolate outliers in individual nodes. CART algorithm works as follow: Constructing the maximum tree which is the most time consuming part then choosing the right tree size and finally performing the classification of new data using a constructed tree [30]. The CART methodology includes automatic class balancing, handles missing values and allows for cost-sensitive learning, dynamic feature construction, and probability tree estimation [14].
- REPTree: Reduced Error Pruning Tree (REPT) is a fast decision tree algorithm. It applies regression tree logic and creates multiple trees in different iterations then finally selects the best one as the final tree. REPTree builds a decision tree based on the information gain and prunes it using reduced error pruning [8]. Pruning techniques have been used to minimize the complexity of tree structure without reducing the accuracy rate of classification. The basic of the REPTree is sorting values for numerical attribute once and handling the missing values by using C4.5's method of using fractional instances[9].

B. Data Description

The dataset is taken from direct marketing campaigns of Portuguese banking institution. It was collected and prepared by S.Moro,R.Laureano and P.Cortez [19], [17]. The dominant marketing campaigns were based on phone calls. The dataset contains (41188) instances and (20) attributes with one output attribute (target). All the available attributes in the dataset and their description are presented in Table 1 [19], [17].

As shown in Table 1, there are three kinds of attributes, which are Categorical, Numerical and Binary. The target attribute (Y) is binary with two classes which are "yes" which indicates that a deposit subscribed by clients and "no" which indicates that no deposit was subscribed by any clients. This dataset has 4640 clients with class label "yes" and 36545 clients with class label "no".

C. Evaluation Measures

A comparison between these algorithms is performed based on some standard performance metrics which are accuracy, precision, True Positive rate (TP) and F-measure based on the confusion matrix of each tree. The confusion matrix is a table that contains a summary of the prediction results of the classification system [31]. A confusion matrix for a binary classifier is shown in Table 2. It includes data about the actual and predicted values obtained by the classification model [24].

True Positive(TP) presents the number of the correct observation assigned to the positive class. True Negative (TN) presents the correct observation assigned to the negative class. The False Positive (FP) is the number of incorrect observations assigned to the positive class. Finally, False Negative (FN) presents the number of incorrect observations assigned by the model to the negative class [31] [3].

A classifier accuracy reflects it's overall prediction correctness and is defined as the number of the correct predictions to the total number of predictions. Accuracy is given by the Formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

True Positive Rate in machine learning referes to sensitivity or recall. It is used to measure the percentage of actual positives which are correctly predicted as positive. Recall is given by the Formula:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Precision is a good measure to determine how precise the model is, and to tell the number of actual positive class among the predicted positive ones. The high precision indicates a small number of FP. Precision is given by the Formula:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

F measure represents both recall and precision with the formula:

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

Moreover, the Receiver Operating Characteristic (ROC) has been considered in the present study as one of the most commonly used metrics to evaluate the performance of classification models. ROC curve is presented by plotting the true positive rate (Y-axis) against the false positive rate (X-axis). An optimal model will have a ROC value of 1.0 [31].

The used dataset considered class-imbalanced data since class "yes" (minority) has very low proportions in

TABLE I. List of attributes and their description.

Attributes	Type	Attributes Descriptions
Age	Numeric	Clients age
Job	Categorical	(”admin.”,”blue-collar”,”entrepreneur”,”housemaid”,”management”,”retired”,”self-employed”,”services”,”student”,”technician”,”unemployed”,”unknown”
Marital	Categorical	(”divorced”,”married”,”single”,”unknown”; note: ”divorced”)
Education	Categorical	(”basic.4y”,”basic.6y”,”basic.9y”,”high.school”,”illiterate”,”professional.course”,”university.degree”,”unknown”)
Housing	Categorical	(”no”,”yes”,”unknown”)
Loan	Categorical	(”no”,”yes”,”unknown”)
Contact	Categorical	(”cellular”,”telephone”)
Month	Categorical	(”jan”, ”feb”, ”mar”, ..., ”nov”, ”dec”)
day_of_week	Categorical	(”mon”,”tue”,”wed”,”thu”,”fri”)
Duration	Numeric	last contact duration in seconds
Campaign	Numeric	number of contacts performed during this campaign and for this client
Pdays	Numeric	number of days that passed by after the client was last contacted from a previous campaign
Previous	Numeric	number of contacts performed before this campaign and for this client
Poutcome	Categorical	(”failure”,”nonexistent”,”success”)
emp.var.rate	Numeric	employment variation rate - quarterly indicator
cons.price.idx	Numeric	consumer price index - monthly indicator
cons.conf.idx	Numeric	consumer confidence index - monthly indicator
euribor3m:	Numeric	euribor 3 month rate - daily indicator
nr.employed:	Numeric	number of employees - quarterly indicator
Y	Binary(categorical)	has the client subscribed a term deposit?(“yes”,“no”)

TABLE II. Confusion Matrix

		Predicted Values	
		Positive(1)	Negative(0)
Actual Values	Positive(1)	***TP***	***FN***
	Negative(0)	***FP***	***TN***

the dataset compared to the class “no” (majority class). Hence, the classifiers are skewed towards the majority class and performed unfairly on the minority class. In such a case, other performance evaluation metrics should be applied in addition to the accuracy. One of the most popular techniques that solve this problem with good evaluating measures is Geometric Mean (G-Mean).

G-Mean is a metric that measures the balance between classification performances on both the majority and minority classes. A low value of G-mean indicates that the positive cases are weakly categorized even if negative cases are correctly classified [1]. G-Mean is given by the equation:

$$G - Mean = \sqrt{sensitivity * specificity} \tag{5}$$

Sensitivity (6) is also called true positive rate or recall. It measures the ratio of actual positives that are correctly classified as positive, while specificity (7) is also called true negative rate that measures the ratio of actual negatives that are correctly classified. .

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

IV. Experiments and Results

This work has used a bank telemarketing dataset from UCI machine learning repository which consists of 41188 instances and 21 attributes collected by [19], [17] then applied four different decision tree algorithms (C4.5, REPTree, RandomTree and SimpleCart). Moreover, dataset is divided using K-Fold cross validation which is one of the most popular methods for evaluating the performance of classification algorithms, especially when the volume of the data set is large [7]. In the Cross validation technique, the data set is divided randomly into K of approximately equal parts(folds). The first fold is used as a testing set, and the remaining K-1 folds are used as training set. This process is repeated K times until each fold has been used as the testing set. Then the model accuracy is calculated as the average of the obtained accuracy in each round [31]. The K value must be chosen wisely. It is usually set to 5 or 10 folds. As K increases, the overlap between training sets also increases. Choosing the value of K equal 10 is more likely and very common to be used because it makes predictions using 90% of the data [25]. Therefore, in this paper the data is split using a 10-fold cross validation to evaluate the predictive model performance.

For easy understanding of the learning process, there is a need to work with an algorithm which gives a maximum classification accuracy rate with simple structure in case of the existence of a huge set of data .

In this experiment, the algorithms are implemented on “Weka”, which is an open-source tool written in java used for data mining tasks. It was developed at the University of Waikato in New Zealand, and it can be executed on many platforms, like Windows, Linux and Macintosh operating systems [28].

Weka provides an easy interface and implementations to different learning algorithms for regression, classification, clustering, association rule mining and attribute selection that can be applied to new datasets [6], [28]. All algorithms import the input file in the form of ARFF format. In this experiment, Windows 10 operating system with 8GB RAM was used to run Weka 3.9.3.

Table 3 presents the experimental results of all the proposed decision tree algorithms applied on the bank dataset. These values represent the rare class “yes”.

Based on accuracy, the following can be observed:

- When applying the SimpleCart algorithm, the correctly classified instances values are proposed directly by Weka software without the need for a manual calculation. 37664 instances from the total 41188 instances are correctly classified, which indicates an accuracy equal to 91.44%. The rate of the true positive values is equal to 0.552, and it has a G-mean with 0.728.
- C4.5 and REPTree algorithms show good accuracy results too with 91.19% and 91.07% values respectively. Additionally, the G-mean values for C4.5 is 0.719 and for REPTree is 0.702.
- When comparing the RandomTree algorithm to

TABLE III. PREDECTIVE METRICS FOR ALL DT ALGORITHMS

Model	Accuracy(%)	TP-Rate/Recall	FP- Rate	Precision	F-Measure	ROC Area	G-Mean
SimpleCart	91.44	0.552	0.040	0.639	0.593	0.903	0.728
C4.5	91.19	0.538	0.041	0.627	0.580	0.884	0.719
REPTree	91.07	0.517	0.039	0.626	0.566	0.903	0.702
RandomTree	88.62	0.475	0.062	0.495	0.484	0.726	0.677

the other used algorithms, it proposed The Correctly Classified Instances value equal to 36499 from the total 41188 instances with a percentage of 88.61%, so it can be seen that it had the least accuracy rate among all other values.

After evaluating the accuracy results, it has been found that SimpleCart and C4.5 have a competitive performance with the highest accuracy of classification compared to the other tress algorithms (REPTree and RandomTree). SimpleCart classified instances 0.25% more accurate than C4.5, which makes SimpleCart algorithm the best model with respect to accuracy.

TP rate and FP rate are also reviewed to compare the results of the different classifiers. The TP rate and FP rate values for SimpleCart are (0.552, 0.040), (0.538, 0.041) for C4.5, (0.517, 0.039) for REPTree and finally (0.475, 0.062) for RandomTree. This shows that SimpleCart has scored the highest TP rate while RandomTree has scored the lowest TP rate. By comparing between the results of the TP rate and FP rate of all the algorithms it is obvious that all these algorithms perform a better prediction for the positive cases. Examining other performance measures, such as the precision and F-Measure of all the algorithms, has showed very close differences in the results. The highest precision value is 0.639, and it is scored by SimpleCart, while RandomTree had the lowest precision value of 0.495. Also Simplecart has scored the highest F-Measure value which is 0.593, while C4.5 , REPTree and RandomTree have scored (0.580, 0.566 and 0.484) respectively.

It is also to be noted that the measurement of the experimental result based on the Receiver Operating Characteristic (ROC) that are also presented in Table 3, shows that SimpleCart and REPTree have an equal values of 0.903. As observed, these values are the highest values among all other tress, followed by C4.5 with a value of 0.884 and RandomTree with a value of 0.726. This indicates that SimpleCart and REPTree predictive models can distinguish between the true positives and negatives with a good result that is nearest to the optimal classification point. Moreover, these models are compared based on G-Mean values that were calculated manually according to the equation (5). SimpleCart has also scored the highest G- Mean value of 0.728 compared to the other trees algorithms.

In addition, during the analysis of these algorithms two parameters are taken into consideration; which are the model construction time and the tree complexity. In terms of complexity, Table 4 presents a comparison between all the proposed trees. The tree complexity is clearly governed by the use of the stop criteria and the pruning process . However, the complexity of the tree is generally measured by the following measurements: the total number of nodes (tree size), the total leaf, the depth of tree, the number of attributes that are used in [16]. As shown in Table 4 , SimpleCart produces 47 total numbers of nodes while REPTree, C4.5 and RandomTree produce 992,1143 and 15505 nodes, respectively. Therefore, SimpleCart is better than all other trees in term of classification accuracy (i.e. the number of instances correctly classified) besides the tree size complexity, which presents an important factor affecting the algorithm efficiency, especially with decision tree classifiers.

Furthermore, the time needed to build the model has been taken into account, As shown in table 4, even though SimpleCart can classify the instances more accurately, it might crash on for larger datasets. Therefore, for large datasets, SimpleCart may be an ineffective algorithm. From the obtained results, the following conclusions are drawn:

- RandomTree is much faster than SimpleCart, as it needs much less running time.
- Although RandomTree does not accurately classify instances as SimpleCart does, it retains larger datasets while SimpleCart crashes.
- Due to RandomTree's ability to handle larger datasets, it can be used for processing unstructured data and for large-scale analysis.

A. Feature Selection and Importance Analysis

After applying all the classification models using all 21 attributes of the analyzed dataset, Weka provides a method for Attribute selection. Attribute selection is the process of removing the irrelevant attributes of the data mining task. Also, it aims to search for a main set of attributes that produce comparable classification results with all used attributes [23]. Even though the accuracy is high, the number of attributes used is relatively high. Hence, Weka is used to reduce the number of attributes to get a relatively better accuracy. Since SimpleCart is the best model according to the performance and tree size, three different selection attributes methods are applied on it, which are:

- InfoGainAttributeEval, which evaluates the relevance of an attribute by measuring the information gain of the attribute with respect to class label[21].
- ChiSquaredAttributeEval, which evaluates the relevance of an attribute by computing the value of the Chi-Squred statistic with respect to the class label[21].
- CorrelationAttributeEval, which evaluates the relevance of an attribute by measuring Pearson's correlation between it and the class label.

Table 5 demonstrates a comparison between the three different methods to extract the most important attributes

TABLE IV. PERFORMANCE OF TREES IN TERM OF COMPLEXITY

Dataset	Instances	Attributes	Algorithm	Time to Build Model(sec.)	Size of Tree
Portuguese banking institution.	41188	21	SimpleCart	19.14	47
			C4.5	1.56	1143
			REPTree	0.66	992
			RandomTree	0.2	15505

and the obtained performance values that were derived from the confusion matrix except for G-Mean which was calculated manually.

By using the reported results in Table 5, it can be seen that the classification accuracy of the SimpleCart model achieved a highest percentage value of 91.4732% when reducing the number of attributes to 10 and by using ChiSquare selection attributes methods. This study will take into consideration G-Mean metric to evaluate the performance of the SimpleCart when reducing the number of attributes, since the dataset has imbalanced classes and G-mean is the best measurement to rely on when the class distribution is imbalanced. Table 5 shows the highest readings of the G-Mean when using Information Gain and ChiSquare for selecting the top 5 most relevant attributes with equal values of 0.736. Fig. 2 shows a performance comparison among the three different selection methods.

Table 6 presents the top 5 ranked features that were obtained directly from Weka by using *Select attributes* Tab. As can be noticed, Information Gain and ChiSquare selection methods have the same G-Mean value because their results provide the same features but with a different order. Moreover, it can also be observed that the attributes *duration, euribor3m* and *nr.employed* are common in the three selection methods, so they are considered the most important features for SimpleCart model. *duration* Indicates that a long contact with clients (in seconds) can increase the probability of successful deposit campaigns. Next comes *euribor3m*, which is short for Euro Interbank Offered Rate, and it is a very important reference for rates in the European markets. The offered euribor rate is for three months and is updated daily. Finally, an interesting outcome indicates that the number of employees (*nr.employed*) who make the calls and contact the clients has an influence on the probability of subscribing a successful deposit.

However, Information Gain and ChiSqure nominated *cons.price.idx* and *cons.conf.idx* (as a monthly average) attributes, meaning that economic indicators like changes in the price levels and the customer confidence in the current and future economy may lead them to save more than to spend.

It is also found that *Pdays* and *emp.var.rate* are influenced and controlled by the decisions of the bank managers. Hence, it can be seen that managers can increase the deposit rate when considering these variables (i.e the number of days after the last interaction with the customer from a previous campaign and employment variations rate).

B. Extracting Interesting Rules

In this part of the study, experiments extract the most important rules in the previously built tree based models. This step is very important to give an insight for decision makers and to assist them in taking efficient decisions utilizing these extracted features. Features have been reduced from 21 features to 5 features which is almost a 75% reduction, and it has been found that the reduction in the number of attributes has achieved better results as presented in Table 5. Moreover, reducing attributes to 5 simplified the practical use of the Simple Cart model for marketers and managers and enabled them to use it in their marketing campaigns.

The most important extracted rules of the top 5 attributes for Simple Cart model are illustrated in algorithm 1. There are 18 *if... then* statements. Take the first statements for example; they can be explained as follows: "If the quarterly average of the total number of the employees is below 5087, bank managers should consider two important features for best response from clients, which are call duration and the euribor rate. If the call duration with the client is less than 172 seconds then the response of the client for depositing money in the bank will not succeed, while when increasing the call duration from 172 to less than 250 seconds and the euribor3m below 0.71649 then the model predict a successful campaign". But if the bank employee performed a long call with the client for a period of time longer than 250 seconds, then the model predicts a successful response. According to these statements the bank managers should pay attention to these three features to bring high profits.

The second part of the algorithm takes into consideration other important features which are constant confidence index *cons.conf.idx* and constant price index *cons.price.idx* in addition to the call duration and euribor rate features. This can be explained as follows: "Given that in the bank number of employees who makes calls more than 5087 (as a quarterly average of the total number of employees) and the call duration is less than 606.5 seconds and the constant confidence index of the client is above -46.65 then the model will predict unsuccessful response from the client. While a successful response prediction was obtained for a call duration more than 835.5 seconds and a constant price index less than 93.956. The analysis above indicates that using data mining technology in the direct marketing campaigns especially in bank sectors, is valuable and will lead to useful and great profits with high competition."

V. CONCLUSIONS

This paper investigates experimentally four types of tree based classification algorithms for predicting the bank direct marketing campaign performance. The classifiers are: SimpleCart, C4.5, RepTree and Random Tree. This type of classifiers was chosen because of its interpretability, flexibility and prediction power. The results show that the best results were achieved using SimpleCart model with an accuracy of 91.44% a precision of 0.639 % and a recall of 0.552%. Furthermore, a feature analysis study is conducted based on different feature selection methods to gain an insight on which variables have more influence

TABLE V. Comparison between the number of the selected attributes and the performance

Attributes Selection Methods	Accuracy(%)	Recall	FP-Rate	Precision	F-Measure	ROC Area	G-Mean
Best Top 5 Attributes							
Information Gain	91.211	0.567	0.044	0.620	0.592	0.902	0.736
ChiSquare	91.211	0.567	0.044	0.620	0.592	0.902	0.736
Correlation	91.4611	0.519	0.035	0.652	0.578	0.929	0.708
Best Top 10 Attributes							
Information Gain	91.4708	0.546	0.038	0.643	0.590	0.908	0.725
ChiSquare	91.4732	0.546	0.038	0.643	0.591	0.906	0.725
Correlation	91.4465	0.511	0.034	0.654	0.574	0.922	0.703
Best Top 15 Attributes							
Information Gain	91.3761	0.542	0.039	0.638	0.586	0.900	0.722
ChiSquare	91.3761	0.542	0.039	0.638	0.586	0.900	0.722
Correlation	91.4198	0.545	0.039	0.640	0.589	0.903	0.724

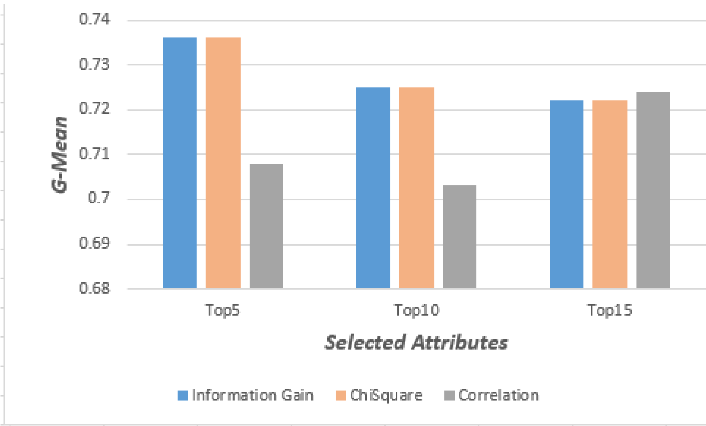


Fig. 2. Comparison

TABLE VI. Top 5 Ranked Attributes

Attributes Selection Methods	Top 5 Ranked Attributes	
Information Gain	0.1094127	duration
	0.1025698	euribor3m
	0.0980036	cons.price.idx
	0.0976254	cons.conf.idx
	0.0896296	nr.employed
ChiSquare	7896.1793	euribor3m
	7618.4243	cons.price.idx
	7570.0707	cons.conf.idx
	7518.7686	duration
	6955.9686	nr.employed
Correlation	0.405	duration
	0.355	nr.employed
	0.325	pdays
	0.308	euribor3m
	0.298	emp.var.rate

in the investigated problem. Best results were gained using top 5 selected features. This analysis showed that the most influencing features are call duration, offered interest rate, number of employees, changes in the prices levels and customer confidence. Such information can be very useful to decision makers, as it can enhance direct marketing campaign, increase the number of clients who subscribe the deposit and lead to a better management of the available resources by focusing on these most influential features. As future work other session features that had not been discussed in the study and may affect the Direct Marketing success can be addressed. Furthermore, this study's results can be evaluated against other sectors. In addition, future work can discuss the effect of these features on different customer segments or investigate different marketing channels rather than phone calls.

References

[1] Josephine Akosa. Predictive accuracy: a misleading performance measure for highly imbalanced data. In *Proceedings of the SAS Global Forum*, pages 2–5, 2017.

[2] Anokye Acheampong Amponsah and Kwaku Agyepong Pabbi. Enhancing direct marketing using data mining: A case of yaa

Algorithm 1: The extracted decision rules

```
if nr.employed < 5087.65 then
    if duration <172.5 then
        no
    else
        if duration<250.5 then
            if euribor3m <0.71649 then
                yes
            else
                if duration <202.5 then
                    no
                else
                    if euribor3m <1.0470 then
                        if duration <234.5 then
                            yes
                        else
                            no
                        end
                    else
                        no
                    end
                end
            end
        else
            yes
        end
    end
else
    if duration <606.5 then
        if cons.conf.idx < -46.6500 then
            if euribor3m <1.5255 then
                if euribor3m <1.3685 then
                    if duration<223.0 then
                        no
                    else
                        yes
                    end
                else
                    no
                end
            else
                if duration 173.5< then
                    no
                else
                    yes
                end
            end
        else
            no
        end
    else
        if duration <835.5 then
            if euribor3m <1.402499 then
                yes
            else
                no
            end
        else
            if cons.price.idx <93.956 then
                yes
            else
                if duration <1614 then
                    if euribor3m < 4.862 then
                        no
                    else
                        yes
                    end
                else
                    yes
                end
            end
        end
    end
end
```

asantewaa rural bank ltd. in ghana. *International Journal of Computer Applications*, 153(7), 2016.

[3] Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, 3(10), 2013.

[4] Hany A Elsalamony. Bank direct marketing analysis of data mining techniques. *International Journal of Computer Applications*, 85(7):12–22, 2014.

[5] Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora, and Naresh Dhami. Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications*, 163(8):15–19, 2017.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[7] Yoonsuh Jung and Jianhua Hu. Ak-fold averaging cross-validation procedure. *Journal of nonparametric statistics*, 27(2):167–179, 2015.

[8] Sushilkumar Kalmegh. Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. *International Journal of Innovative Science, Engineering & Technology*, 2(2):438–446, 2015.

[9] Sushilkumar Rameshpant Kalmegh. Comparative analysis of weka data mining algorithm randomforest, randomtree and ladtree for classification of indigenous news data. *International Journal of Emerging Technology and Advanced Engineering*, 5(1):507–517, 2015.

[10] D Lavanya and Dr K Usha Rani. Analysis of feature selection with classification: Breast cancer datasets. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(5):756–763, 2011.

[11] Wei Li, Xuemei Wu, Yayun Sun, and Quanju Zhang. Credit card customer segmentation and target marketing based on data mining. In *2010 International Conference on Computational Intelligence and Security*, pages 73–76. IEEE, 2010.

[12] Ying Li, Pavankumar Murali, Nan Shao, and Anshul Sheopuri. Applying data mining techniques to direct marketing: Challenges and solutions. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 319–327. IEEE, 2015.

[13] Charles X Ling and Chenghui Li. Data mining for direct marketing: Problems and solutions. In *Kdd*, volume 98, pages 73–79, 1998.

[14] Venkatadri Marriboyina. A comparative study on decision tree classification algorithms in data mining. 04 2018.

[15] Merve Mitik, Ozan Korkmaz, Pinar Karagoz, Ismail Hakki Toroslu, and Ferhat Yucel. Data mining based product marketing technique for banking products. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 552–559. IEEE, 2016.

[16] W Nor Haizan W Mohamed, Mohd Najib Mohd Salleh, and Abdul Halim Omar. A comparative study of reduced error pruning method in decision tree algorithms. In *2012 IEEE International conference on control system, computing and engineering*, pages 392–397. IEEE, 2012.

[17] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[18] Sérgio Moro, Paulo Cortez, and Paulo Rita. A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. *Expert Systems*, 35(3):e12253, 2018.

[19] Sergio Moro, Raul Laureano, and Paulo Cortez. Using data mining for bank direct marketing: An application of the crisp-dm methodology. In *Proceedings of European Simulation and Modelling Conference-ESM'2011*, pages 117–121. EUROSIS-ETI, 2011.

[20] Sérgio Moro, Raul Laureano, and Paulo Cortez. Enhancing bank direct marketing through data mining. In *Proceedings of the 41th European Marketing Academy Conference (EMAC)*. European Marketing Academy, 2012.

[21] Jasmina Novaković. Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*, 21(1), 2016.

[22] Youqin Pan and Zaiyong Tang. Ensemble methods in bank direct marketing. In *2014 11th International Conference on Service Systems and Service Management (ICSSSM)*, pages 1–5. IEEE, 2014.

[23] Nikita Patel and Saurabh Upadhyay. Study of various decision tree pruning methods with their empirical comparison in weka. *International journal of computer applications*, 60(12), 2012.

[24] Tina R Patil, SS Sherekar, et al. Performance analysis of naive bayes and j48 classification algorithm for data classification. *International journal of computer science and applications*, 6(2):256–261, 2013.

[25] Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. *Encyclopedia of database systems*, pages 532–538, 2009.

[26] Philip J Sallis, William Claster, and Sergio Hernández. A machine-learning algorithm for wind gust prediction. *Computers & geosciences*, 37(9):1337–1344, 2011.

[27] Shiju Sathyadevan and Remya R Nair. Comparative analysis of decision tree algorithms: Id3, c4. 5 and random forest. In *Computational Intelligence in Data Mining-Volume 1*, pages 549–562. Springer, 2015.

[28] Trilok Chand Sharma and Manoj Jain. Weka approach for comparative study of classification algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(4):1925–1931, 2013.

[29] Lilian Sing'oei and Jiayang Wang. Data mining framework for direct marketing: A case study of bank marketing. *International Journal of Computer Science Issues (IJCSI)*, 10(2 Part 2):198, 2013.

[30] Roman Timofeev. Classification and regression trees (cart) theory and applications. *Humboldt University, Berlin*, 2004.

[31] Efraim Turban, Ramesh Sharda, and Dursun Delen. *Business intelligence and analytics: systems for decision support*. Pearson Higher Ed, 2014.

[32] Chakarin Vajiramedhin and Anirut Suebsing. Feature selection with data balancing for prediction of bank telemarketing. *Applied Mathematical Sciences*, 8(114):5667–5672, 2014.

[33] K Wisaeng. A comparison of different classification techniques for bank direct marketing. *International Journal of Soft Computing and Engineering (IJSCE)*, 3(4):116–119, 2013.

Modelling an Indoor Crowd Monitoring System based on RSSI-based Distance

Syifaul Fuada[1]
Program Studi Sistem Telekomunikasi
Universitas Pendidikan Indonesia
Bandung, Indonesia

Trio Adiono[2]
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia

Prasetiyo[3]
School of Electrical Engineering
Korea Advanced Institute of Science and Technology
Daejeon, South Korea

Hartian Widhanto Shorful Islam[4]
Stream Intelligence Ltd.
London, United Kingdom

***Abstract*—This paper reports a real-time localization algorithm system that has a main function to determine the location of devices accurately. The model can locate the smartphone position passively (which do not need a set on a smartphone) as long as the Wi-Fi is turned on. The algorithm uses Intersection Density, and the Nonlinear Least Square Algorithm (NLS) method that utilizes the Lavenberg-Marquart method. To minimize the localization error, Kalman Filter (KF) is used. The algorithm is computed under Matlab approach. The most obtained model will be implemented in this Wi-Fi tracker system using RSSI-based distance for indoor crowd monitoring. According to the experiment result, KF can improve Hit ratio of 81.15 %. Hit ratio is predicting results of a location that is less than 5 m from the actual area (location). It can be obtained from several RSSI scans, the calculation is as follows: the number of non-error results divided by the number of RSSI scans and multiplied by 100%.**

Keywords—Wi-Fi tracker system; RSSI-based distance; intersection density method; Nonlinear Least Square (NLS) method; Kalman Filter (KF)

I. Introduction

In 4.0 era, crowd monitoring/tracking system has become very useful application because it provides some summaries and insights about flow, direction, density and activity of people in certain public and private areas. A prior work described the use of a real-time Wi-Fi tracking system for business intelligence in the retail company [1]. The methods can be used in the crowd monitoring system, e.g., image-based and non-image-based method. Generally, image or video-based system requires high-cost and complex computation. Also, that approach has several other disadvantages: it only covers a small line-of-sight (LoS) area and difficult to obtain high estimation accuracy when overlap and occlusion exist in the crowd. In other hands, the video-based method does not work in dark or smoke environments and also less privacy of the target [2]. The non-image-based method can overcome drawbacks in image-based methods, especially in the cost factor, and it can cover a high LoS area.

Nowadays, most people bring their smartphones everywhere they go. The Wi-Fi access point facilities have also been installed in many places; it will trigger most people to turn on their Wi-Fi on smartphones. We can use this fact to track-down them using a Wi-Fi-based approach. The Wi-Fi devices on a smartphone will reveal their MAC address data through probe-request data whenever the Wi-Fi devices on. In other words, when the device activates Wi-Fi, they will eventually broadcast probe request data containing useful information; for example, MAC addresses data of device and time-stamp [3-5]. Because the MAC address is a unique identifier, it can be used to identify the presence of people in a particular location. In this work, we used Wi-Fi-based RSSI localization to estimate or monitor the crowd pattern at certain places. A similar approach has been proposed in work at [6] and [7].

Radio Frequency (RF) technologies such as GPS, RFID, Bluetooth, and Wi-Fi, use radio signals to find the device's location. Generally, we put the sensor (node) to sense the RF signal parameter then estimate the location from that parameter. Some of RF parameter that has been used by researchers: Time difference of Arrival [8], Time of Arrival [9], Angle of Arrival [10], and RSSI [11]. The RSSI-based localization provides a simpler node compared to the other methods. But this approach gives a major problem to the detection accuracy due to the high variance on RSSI value, the techniques to overcome it, e.g., Bayes Filter [12], Particle Filter, and so on.

There is a various method that has been proposed in the Wi-Fi RSSI-based indoor localization, like Finger-printing, Distance-based localization, etc. This work will provide GUI to monitor the indoor crowd, based on the RSSI localization method. However, the biggest problem in the real application is, the RSSI value measured at the node is unstable. They keep changing dynamically because of the presence of noise. The distribution of RSSI value most likely to be Gaussian. Therefore, correction values by using the Kalman Filter (KF) is made.

In this work, the set of RSSI data is scanned in multiple sensors (nodes). One of two recommended localization algorithms: Intersection Density and Non-Linear Least Square (NLS) will be used in this experiment. The localization algorithm is used to estimate the location of each device precisely. Whereas the KF algorithm is used to overcome the noises in RSSI signals that are scanned in sensor.

This paper is composed by four sections: 1) Introduction that discusses why RSSI-based localization method is selected for Wi-Fi tracking application, 2) Methods discusses proposed system and the used algorithms, 3) Results and Analysis, and the last section is 4) Conclusion.

II. Methods

A. Proposed Systems

Fig. 1 shows a proposed architecture of the Wi-Fi Tracker system. It consists of two main parts: the node system and the server system. The node system has a primary function to sniff the devices (smartphones) data on the site. Then the server will compute and analyze the data and make a summary about the situation on the dashboard. As informed by [3-5], when the devices activate their Wi-Fi, they will broadcast wireless signature data containing unique information about the devices, which is a MAC address. According to our previous experiment, as in [5], the packet request data is Wi-Fi's packet data associated with it. From that data, we also got the information besides MAC address data: Timestamp and RSSI.

In the first step, we placed the node at several points on the site. Afterward, the devices on the sites will broadcast the packet request data to the surrounding access point (AP). The smartphone broadcasts the packet request data when its Wi-Fi is turned on, and they do not have to be connected with Node. In other cases, the smartphone can be connected with the surrounding AP. As long as they broadcast the packet request data, the system will continue work. Every node on the sites will sniff the packet data emitted by the devices. The node must be in monitoring mode to sniff those packet data. Then the nodes will send the data to the server.

Before the data sent to the server, the data will be encrypted using TLS/SSL with 1024-RSA encryption. The nodes will connect to our proposed access point (Wi-Fi tracker access point) and send the data to the server using Message Queuing Telemetry Transport (MQTT) protocol.

The server will collect the data from all nodes by subscribing to the MQTT broker based on the designed topic. The server will decrypt the packet data, then collect the data and organize them based on the MAC address information. Therefore, in the server, we will get a set of MAC address and their corresponding RSSI in every node. From this set of data, we will compute the location of each device using the algorithm (NLS and KF algorithms). The raw data and processed data (smartphone and its location) will be stored in a database. In our system, we used MongoDB as a database system. We will also provide the dashboard based on a web application that runs on our server. Beside serve basic configuration to the system, the web will also provide the information (analyzed data).

But in this work focuses on the algorithm part. RSSI data is collected and then computed through Matlab simulation. The most optimum algorithm is then implemented in a server using the Phyton script (further research). Thus, our system will have a high-accuracy in tracking the devices. Three approaches: Intersection Density, NLS, and linear KF are computed.

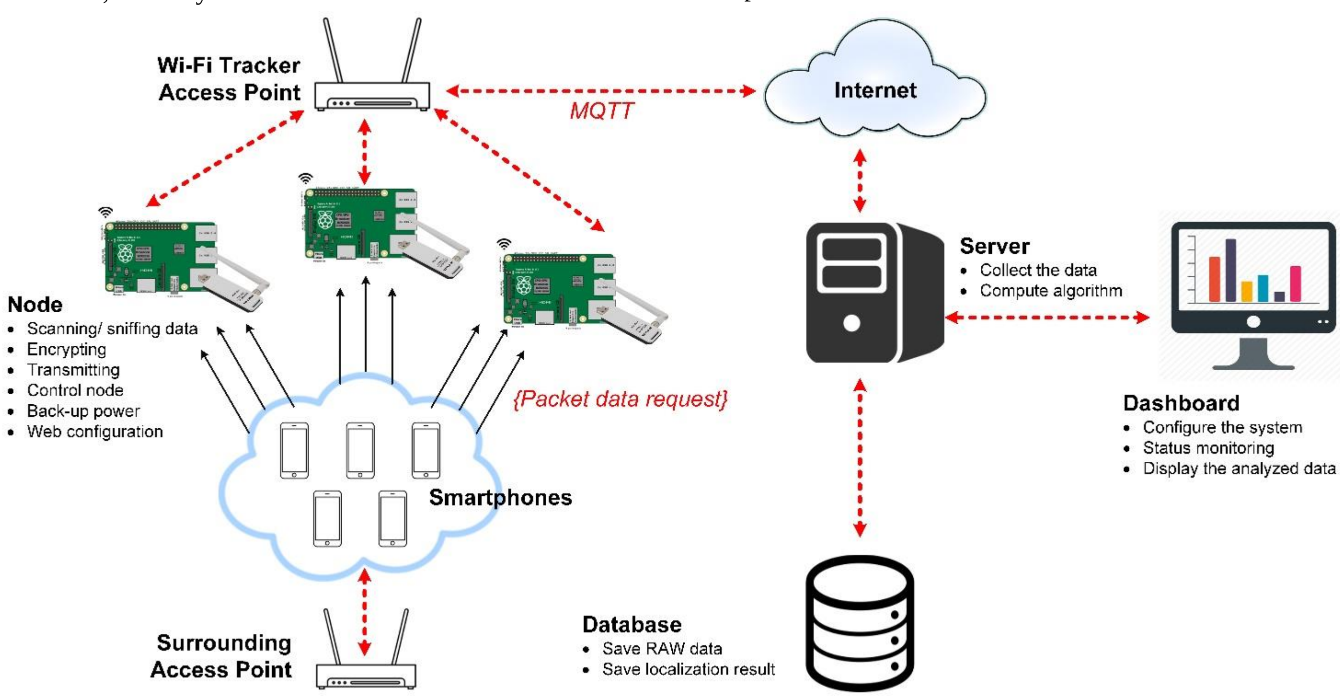


Fig. 1. Wi-Fi Tracker System Architecture.

B. Localization Algorithm

A node network with a known location is expected to get the RSSI value from the devices. Then from this set of RSSI measurements, the device location can be predicted by estimating the distance between the smartphone and the node. First, we estimate the distance between nodes and devices using the path-loss model and RSSI values. We can formulate a mathematical representation of our problem as follows:

- Given m known nodes, each is located at (x_i, y_i) $i = 1, 2, \ldots, m$. The RSSI value measured at node i-th are P_i $i = 1, 2, \ldots, m$
- From the data set of $\{(x_i, y_i, P_i)\}$ then we estimate the location of devices (x, y)

There are two general steps of RSSI-based distance estimation, first is Distance estimation, by using path-loss model, the value of P_i is used to estimate the distance between the smartphone and node i-th (d_i), and the second one is (b) Location estimation. From the set of measurable distance $\{d_i\}$, the smartphone location (x, y) is estimated. This step can be done using localization algorithm, such as Triangulation, Trilateration, Intersection Density, Linear Least Square, NLS, etc. In this work, we have tried two methods: Intersection Density and NLS. We compare them which is the best one to be used as the localization algorithm.

- Path-Loss model

Path-loss is a reduction in power density of the electromagnetic waves as it propagates through space (attenuation). It represents signal level attenuation caused by free-space propagation, reflection, diffraction, absorption, and scattering. There are various path-loss models that have been established to represent Wi-Fi communication as well as the condition in the room. In this work, we used the log-distance path-loss model as in Eq.1,

$$P(d) = P(d0) - 10\,\alpha\, log\frac{d}{d0} + X_\sigma \tag{1}$$

Where $P(d)$ is power at distance d, $P(d0)$ is reference power that emitted by smartphone (power at distance d_0 = 1m), α is Path-loss exponent (depending on the surrounding environment, related to attenuation factor). Later, X_σ is Zero-mean, and σ-variance is a random variable (from noise, shadowing, multi-path effect).

The path-loss exponent value depends on the surrounding environment, thus to get the precise value of this parameter, the calibration must be performed. The path-loss exponent value can be determined by measuring the RSSI value for several minutes at a specific distance in the room. From that data, by using the path-loss model in Eq. 1, the value of α can be computed. Table I shows path-loss exponent.

The reference power value emitted by a smartphone has different from other smartphones. Due to our application is intended to detect many smartphone types at once, this parameter value cannot be determined by doing a calibration only. Hence, this is another point that we have to consider when performing the localization algorithm.

TABLE. I. PATH-LOSS EXPONENT VALUE

Environment	Path-Loss Exponent
Free Space	2
Urban area cellular radio	2.7 - 3.5
Shadowed urban cellular radio	3.0 - 3.5
In building LOS	1.6 to 1.8
Obstructed in building	4 to 6
Obstructed in factories	2 to 3

When a Wi-Fi signal encounters another medium with different electrical properties, there is partly reflected signal and partly absorbed signal. The reflection coefficient is a complex function of the material properties and generally depends on signal frequency, polarization, and angle of incidence. The previous model (Eq. 1) is suitable for Line-of-Sight (LOS) indoor environment. For the Non-Line-of-Sight (NLOS) environment, there is a more precise model that considers the attenuation factor from the existing obstacles. The multi-wall model can be expressed in Eq. 2,

$$P(d) = P(d0) - 10\,\alpha\, log\frac{d}{d0} + X_\sigma + \sum_{i=1}^{k} A_i \tag{2}$$

Where k is a number of obstacles between transmitter and receiver and A_i is an attenuation factor for obstacle *i-th*. Several examples of attenuation factor value for 2.4 GHz Wi-Fi signal in various materials can be seen in Table II.

Actually, in our application, this factor can only be controlled by setting the path-loss exponent value through a calibration in the room. Several precise path-loss models can represent the actual situation in the real case. Still, we use a simple path-loss model for this work. In a further improvement, it can be used for the more precise model that has higher accuracy to represent path-loss in the room.

- Intersection Density

The Intersection Density method estimates the smartphone location by utilizing pairs of known node locations to generate circles. By using multiple different pairs of the node, multiple circles can be derived, each of which intersects at smartphone location in the absence of noise and measurement errors. Of course, noise and measurement errors always exist in our measurements. This fact causes the intersection point does not intersect in a single point. However, Intersection Density assumes that a number of the intersection will be the highest in the surrounding of a smartphone location. Therefore, the smartphone location will be determined in the area that has the most intersections.

TABLE. II. ATTENUATION FACTOR FROM SEVERAL OBSTACLES

Obstacle	Attenuation factor
Concrete wall 10 cm	-16 dBm
Concrete wall 20 cm	-29 dBm
Metal	-26 dBm
Glass wall	-6 dBm
Wooden wall	-4 dBm
Human body	-3 dBm

First, from the log-distance model, by assuming that the surrounding has the same path exponent value for all directions, then we can obtain power difference as Eq. 3,

$$P_{ij} = P_i - P_j = 10\,\alpha \log \frac{dj}{di} \quad (3)$$

Next step we define and measure the distance ratio for all known nodes as expressed by Eq. 4,

$$d_{ij} = \frac{d_i}{d_j} = 10^{\frac{P_j - P_i}{10\alpha}} \quad (4)$$

The Intersection Density algorithm will map the set of node location $\{(x_i, y_i)\}$ and distance ratio $\{d_{ij}\}$ into set of circle with center $\{C_{ij}\}$ and radius $\{r_{ij}\}$ as Eq. 5 and Eq. 6 respectively,

$$C_{ij} = \left(\frac{x_j d_{ij}^2 - x_i}{d_{ij}^2 - 1}, \frac{y_j d_{ij}^2 - y_i}{d_{ij}^2 - 1}\right) \quad (5)$$

$$r_{ij} = \sqrt{\left(\frac{x_j d_{ij}^2 - x_i}{d_{ij}^2 - 1}\right)^2 + \left(\frac{y_j d_{ij}^2 - y_i}{d_{ij}^2 - 1}\right)^2 - \frac{d_{ij}^2 x_j^2 + d_{ij}^2 y_j^2 - x_i^2 - y_i^2}{d_{ij}^2 - 1}} \quad (6)$$

The next step is to find location where the circles intersect each other. This can be done by generate circle equation, then find the location of the intersection by solving the equation for each pair of circles. Then we divide the map into several grid areas. Later, the smartphone location (x, y) can be estimated in the location that has the most intersection points [13]. For example, in Fig. 2, the location of the smartphone can be estimated in the grid with the x mark where on the grid there are the highest intersection point.

- Non-Liniear Least Square Algorithm (NLS)

This method finds the smartphone location by forming the objective function that represents the mean square error between the measurement and model. Then by using optimization function, we estimate the smartphone location (x, y) that minimizes our objection function. Because our objective function is nonlinear with the respects of the variable (x and y), then we called it NLS. From the path-loss model, we can calculate the power difference as Eq. 7,

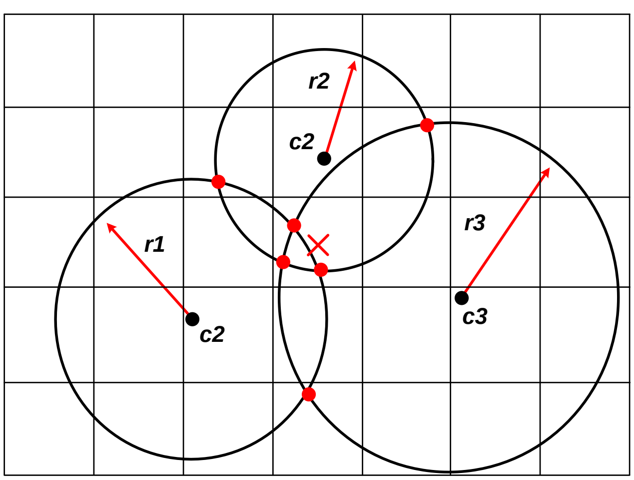


Fig. 2. Illustration of Intersection Density Method.

$$P_{ij} = P_i - P_j = 10\,\alpha \log \frac{dj}{di}$$

$$= 10\,\alpha \log \frac{\sqrt{(x-x_j)^2 + (y-y_j)^2}}{\sqrt{(x-x_i)^2 + (y-y_i)^2}} = 5\alpha \log \frac{(x-x_j)^2 + (y-y_j)^2}{(x-x_i)^2 + (y-y_i)^2} \quad (7)$$

This power difference equation is used because we want to eliminate the power reference that emitted by smartphone, which are unknown for us. Still, this can be done by assuming that path-loss exponent is equal in all rooms. From the measured RSSI in nodes, we can compute the measured power different $\overline{P_{ij}}$. Afterward, we correct the results using KF and then define objective function as the sum of squares of differences between the measured value and theoretical value as Eq. 8,

$$Q(x, y) = \sum_{i<j} \left[\overline{P_{ij}} - 5\alpha \log \frac{(x-x_j)^2 + (y-y_j)^2}{(x-x_i)^2 + (y-y_i)^2}\right]^2, \; 1 \le i < j \le m \quad (8)$$

NLS algorithm finds the value of (x, y) that minimizes the objective function of $Q(x, y)$. Some common methods that can be used to solve the NLS problem, i.e., Gradient descent method, Gauss-Newton method, and Levenberg-Marquardt method. But, in this work, to find the solution, we used the Lavenberg-Marquart method.

The objective function $Q(x, y)$ is nonlinear. In case, we can linearize the objective function using the Taylor series expansion; this method transforms the NLS into Linear Least Square method. We did not try it yet, however we think that the nonlinear model is more suitable for our systems, so we will use NLS instead. Eq. 9 shows objective function that consider initial power device,

$$Q(x, y) = \sum_{1 \le j \le m} [P_{init} - 5\alpha \log\left(x - x_j\right)^2 + \left(y - y_j\right)^2 - \overline{P_j}]^2 \;, -50\,dBm < P_{init} < -30\,dBm \quad (9)$$

C. Kalman Filter

If we observe the RSSI value that is scanned in node, we will find that the value is continuously changing even for the smartphone placed in the same location. There are so many factors that cause it. But in the real case, that is what actually happened; it will decrease the accuracy of our algorithm. To suppress this problem, we used KF to reduce the noise that happens in the node when performing a measurement.

KF works well for the systems which are continuously changing; it can predict uncertain information about a dynamic system, what the system is going to next, and its value. This filter has some advantages: it requires less memory (only needs a previous state value other than the whole history) and also has fast computation (suitable for the real-time applications).

Generally, KF has two steps, which is the prediction step and the correction step. In the prediction step, KF makes a prediction based on the previous state. Then in the correction step, KF will correct the prediction value with regard to the measurement in this state. The general KF problem is stated as Eq. 10 and Eq. 11,

$$X_{k+1} = f(X_k) + w_k \quad (10)$$

$$Z_k = h(X_k) + v_k \quad (11)$$

Where X is system state vector, *f(.)* is transition function, Z is measurement vector, *h(.)* is measurement function, w is process noise, and v is measurement noise. Both w and v are zero mean Gaussian distribution with covariance Q and R, respectively.

D. Linear Kalman Filter Calculation

In Linear KF, both the transition function *f(.)* and the measurement function *h(.)* are linear functions. We can divide into two steps as follows:

- Prediction step: In this step, we have to predict the value based on the previous state. Project the state ahead using Eq. 12, while project the covariance matrix using Eq. 13,

$$X_{k|k-1} = F * X_{k-1} \quad (12)$$

$$P_{k|k-1} = F * P_{k-1} * F^T + Q \quad (13)$$

Where F is a transition matrix model, and the covariance matrix P represents the level of certainty of our prediction value.

- Correction step: In this step, we correct our prediction based on the measurement result at this time. First, we have to compute the Kalman gain as expressed by Eq. 14,

$$K_k = P_{k|k-1} * H^T * (H * P_{k|k-1} * H^T + R)^{-1} \quad (14)$$

Then, correct the prediction based on the measurement as Eq. 15,

$$X_k = X_{k|k-1} + K_k * (Z_k - H * X_{k|k-1}) \quad (15)$$

Additionally, we have to update the covariance matrix as Eq. 16,

$$P_k = (I - K_k * H) * P_{k|k-1} \quad (16)$$

Where H is observation matrix model.

In this application, each time step, we collect the RSSI value from all nodes then uses KF for each before using them for our algorithm. Fig. 3 illustrates how the KF is used in this work; the KF works for each node. Each computation requires the previous step value so that each time the calculation has been made. The value must be stored in memory. We only need previous value, so the other values can be removed after the computation has been done. KF block diagram can be seen in Fig. 4.

Afterward, we define our state model which represent the RSSI value and the velocity of smartphone movement. This variable is intended to predict the RSSI when the smartphone is moved or not. We define our state vector as two variables: RSSI – velocity $\dot{x}$ (Eq. 17), and observation vector as the result of scanned RSSI from node (Eq. 18),

$$X = \begin{bmatrix} RSSi \\ \dot{x} \end{bmatrix} \quad (17)$$

$$Z = [\,RSSI\,] \quad (18)$$

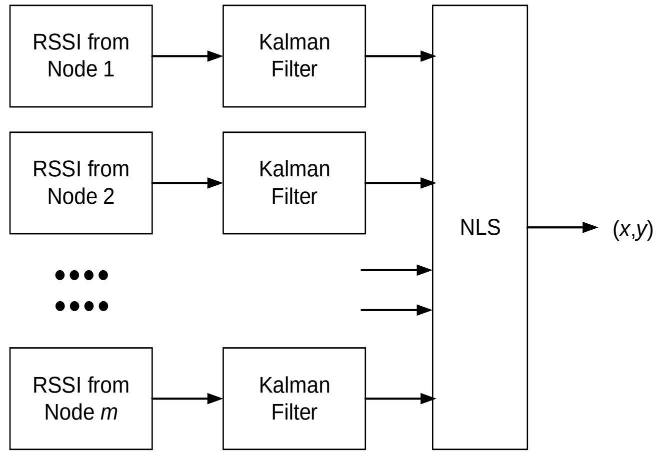


Fig. 3. Algorithm Architecture.

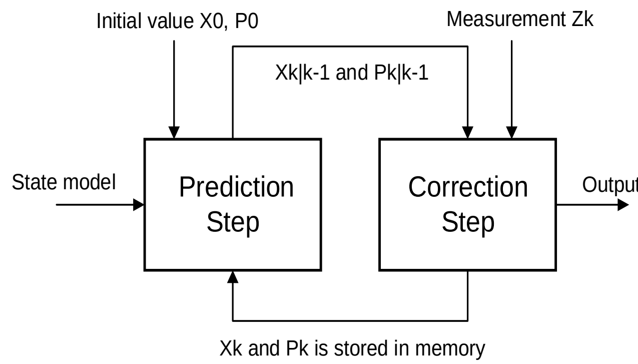


Fig. 4. Steps of KF.

Then KF model used in this work can be expressed as Eq. 19 to Eq. 22, where *dt* is time interval between the current step and the previous step.

$$F = \begin{bmatrix} 1 & -8 * dt \\ 0 & 1 \end{bmatrix} \quad (19)$$

$$H = [1 \quad 0] \quad (20)$$

$$Q = \sigma^2 \begin{bmatrix} dt^3/3 & dt^2/2 \\ dt^2/2 & dt \end{bmatrix} \quad (21)$$

$$R = \left[\left(\frac{3.0 * RSSI + 340}{70}\right)^2 \right] \quad (22)$$

III. Results and Discussion

A. Path-Loss Simulation

We first examine the path-loss exponent value in the room to get a suitable path-loss model. We measure the RSSI value based on two parameters, i.e., several times against several known distances in the room. Fig. 5 visualizes a sample of obtained RSSI from the observed device. The measurement method is elaborated in [5].

By using the path-loss model as Eq. 1, we can compute the path-loss exponent value for each measurement as shown in Fig. 6. We can estimate the path-loss exponent value of the room to be 2.4 – 2.7. We compute it in the room based on the selected room reference that is Research and Community Service (CRCS) Institut Teknologi Bandung building 1st floor.

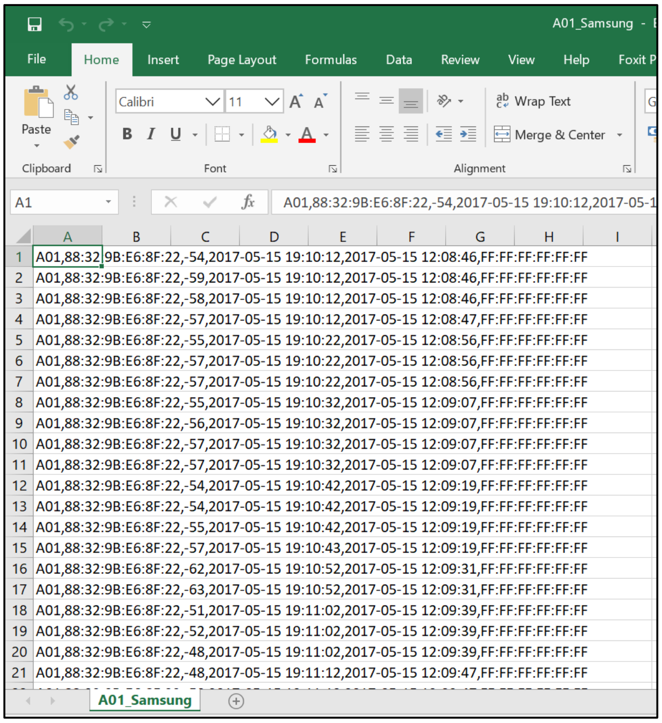


Fig. 5. Example of RSSI Data Logger from "X Smartphone".

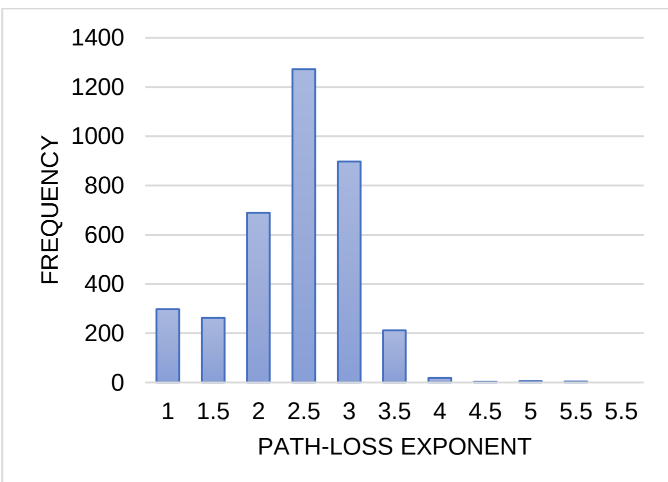


Fig. 6. Path-Loss Exponent Distribution.

B. Intersection Density Simulation

In second step, we measure the Intersection Density algorithm performance. At this simulation test, we take non-real-time data from each node. We get RSSI data for 5 minutes for each location, as illustrated in Fig. 5, then we take its average to represent the scanned RSSI value on each node. This data will be used in the Intersection Density algorithm. The computation is done using Matlab (offline computation). Suppose "X smartphone" and "Y smartphone" as detected devices.

Fig. 7 illustrates several results of the Intersection Density algorithm. The real-location smartphone is located in a green mark. The estimated location lays in the area that has the highest intersection point.

When the scanned RSSI does not represent the real power, the intersection point will be parted to each other. In this case, it is difficult to estimate the location because the intersection area will be "large". Then if we used the real-time data, most of them do not intersect in our valid area. Hence, we assumed that this algorithm has low accuracy for real-time data, and we decide to try to use different methods, i.e., using NLS approach.

C. NLS Simulation

In this test, we measure the performance in our algorithm. We use real-time data that are measured in our server, although the computation is done in offline mode. Fig. 8 depicts the simulation result of data collected by our server for 5 minutes. The "X smartphone" is placed in the center of the room (pointed by a red mark). And four nodes are placed in the corner of the room (points A, B, C, and D). The blue circle represents the estimated point of the smartphone location.

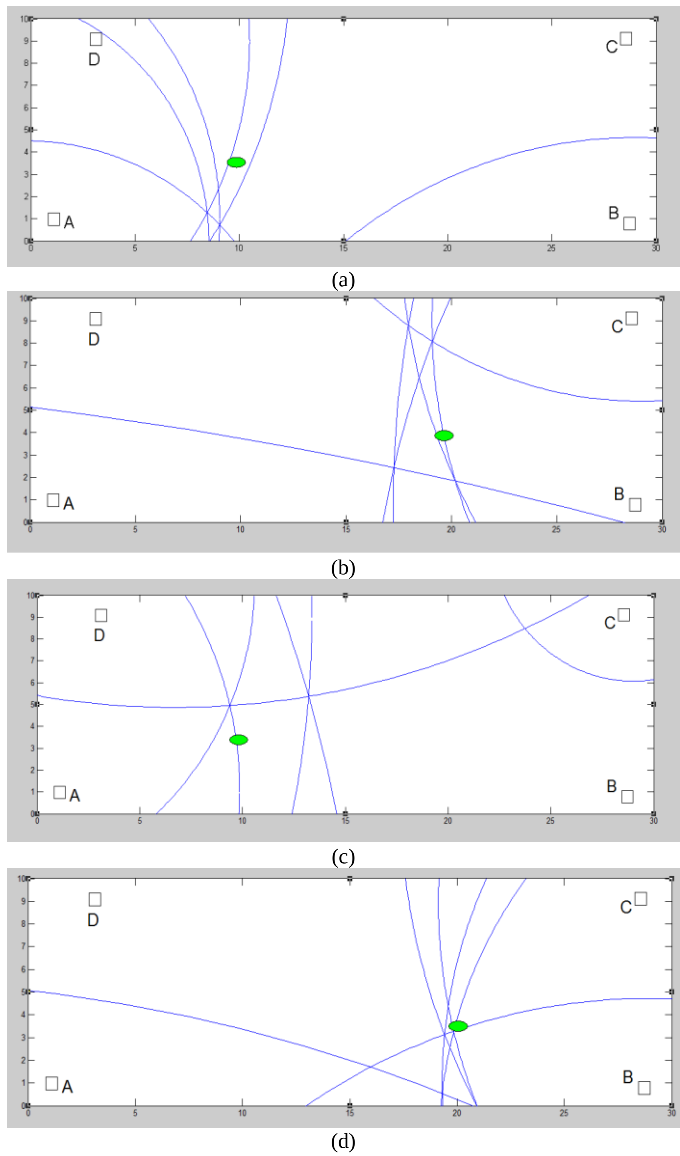


Fig. 7. Intersection Density Result: (a) X Smartphone Location I (b); X Smartphone Location II; (c) Y Smartphone Location I; (d) Y Smartphone Location II.

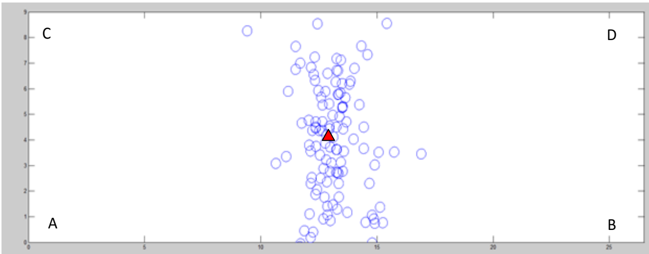


Fig. 8. NLS Simulation Result (X Smartphone) –Location I, using Eq. 8 Approach.

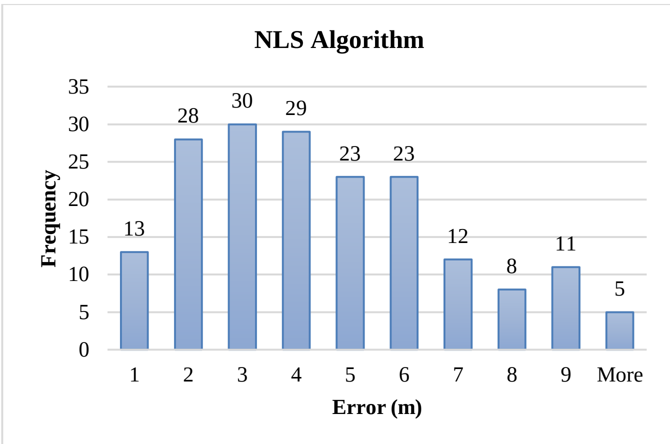


Fig. 9. Error Measurement of NLS Algorithm–Location I, using Eq. 8 Approach.

Then we measure the error by comparing the result as obtained in Fig. 8 to the real-location. If we define the hit ratio as the number when the estimated location has error 5 m or less in several periods, then this measurement has hit ratio = 67.5 %. Fig. 9 shows the obtained graph that represents the error calculation. For NLS simulation, we only take Fig. 7(a) as a sample.

In the latest observations (real-time test), we found a problem using the above algorithm (Eq. 8). Some devices that are placed outside the room, can also be detected as inside the room. After making some observations and analysis, we found what caused it. In Eq. 8, we tried to eliminate the initial power value that smartphone transmits. We did this because initial power is not a fixed variable that different smartphone gives different value. Let's say that our sensor A, B, C, and D read RSSI value from a smartphone of -50 dBm, -50 dBm, -51 dBm, and -51 dBm, respectively. Then we have the same smartphone, but we place it in different locations, and our sensor A, B, C, and D read RSSI of -60 dBm, -60 dBm, -61 dBm, and -61 dBm respectively. The above algorithm (Eq. 8) will predict the same location for both conditions; this is because they have the same power different.

To overcome this problem, we involve additional cases when using the NLS algorithm. First, we use the objective function as Eq. 9, which considers the initial power of the smartphone to decide whether a smartphone is placed inside or not. We use several initial power values and decide which one is the most optimum by looking at cost function (residual error from objective function in the solution), from that we decide which smartphone is in inside or outside the room. Then if the device is inside, we use the previous objective function (Eq. 8) to get a better estimate location by eliminating initial power value.

D. KF Simulation

As in previous step, we first measure the RSSI value of a node, then we use the KF. Lastly, we integrate it with the NLS algorithm. Fig. 10 visualizes the result comparison of raw data before and after filtering. The data is captured for a specific time, and the interest-device (we only interested in one device) which is moving around the node. The filtering technique is used to remove 'random' signal strength, which captured in a single node. As we observe, raw signal strength (blue-colored line) can be filtered out using KF (red-colored line).

In line to Fig. 10, we can conclude that KF is able to overcome the 'random' signal strength. Later, we apply KF in each node. We compare the result of the algorithm before and after applying the KF as shown in Fig. 11. It illustrates the room with a dimension of 30 m x 10 m (the CRCS ITB room size). While A, B, C, and D point represents node system and placed on each corner. The smartphone is placed in the center of room (red mark). The blue circle indicates the result of the algorithm for several times. Later, we analyze the error of the predicted location with the actual location of the smartphone. Fig. 12 depicts the analysis result. In the accuracy of 5 m, we have a hit ratio of 68.115 %.

Then for the same data, we use KF and the result of the algorithm is shown in Fig. 13, whereas the obtained graph is shown in Fig. 14. Using a filter, we have a hit ratio of 81.15 %. According to the simulation result, we can summarize that KF can improve the hit ratio from 68.115% to 81.15 %.

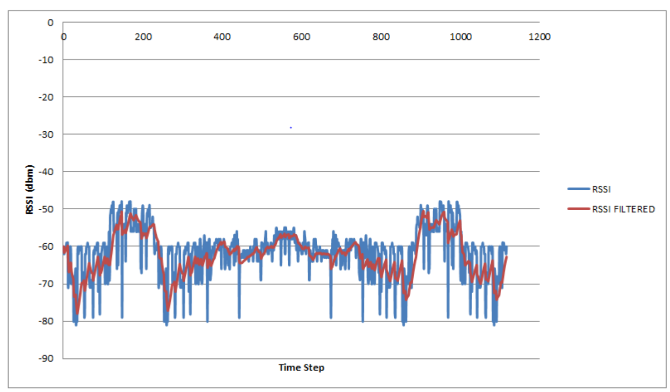


Fig. 10. The Scanned RSSI in the Node before and after using KF.

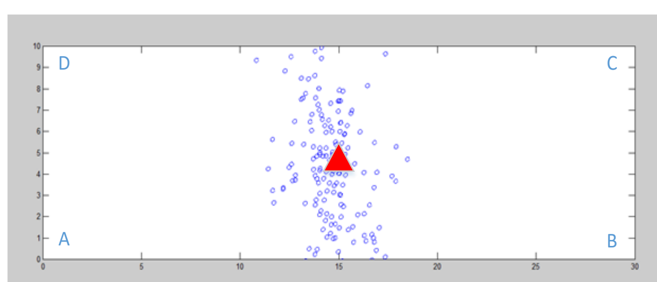


Fig. 11. The Predicted Location without KF, using Eq. 9 Approach.

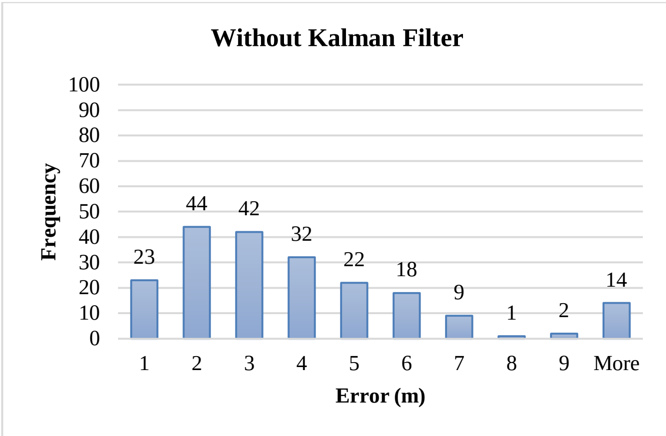


Fig. 12. Error Analysis of the Algorithm without KF, using Eq. 9 Approach.

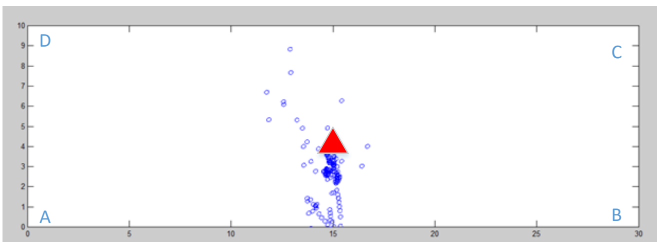


Fig. 13. The Predicted Location using KF.

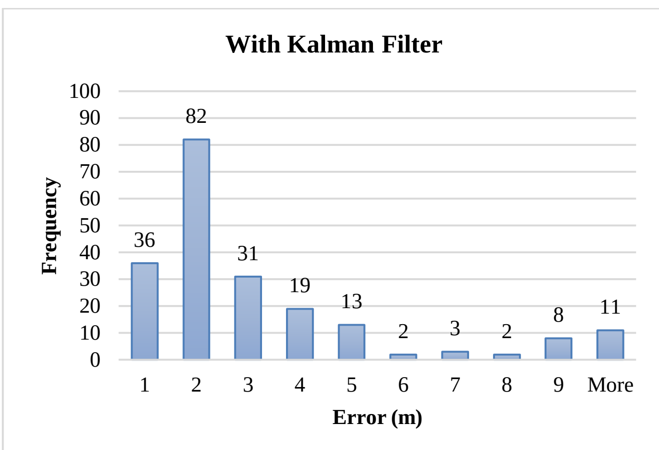


Fig. 14. Error Analysis of the Algorithm with KF.

Based on the obtained information, it is confirmed that the NLS method has a better accuracy compared to the Intersection Density method in a real-time case, and KF can improve the accuracy of the NLS almost 13%. We predict that later, UKF will perform better performance than KF. The most optimum algorithm will be implemented in the RSSI-based Wi-Fi tracking application for indoor environment, such as a system presented by recent works: G. Pipelidis, et al. [14], and Fernandez, et al. [15]. But, our system will has a complete features compared to [14-15].

IV. Conclusion and Future Works

The localization algorithm for the Wi-Fi tracker system has been modeled in this paper. It has the primary function to determine the location of devices (e.g., smartphone, laptop, tablet, etc.) in an indoor environment based on the information which is collected in the server. Two localization algorithm using distance-based methods (i.e., Intersection Density and NLS algorithms) have been tried under a Matlab simulation. The Intersection Density algorithm performs well in non-real time data (with error within 3 – 5 m). But it has a low-accuracy for real-time data (with error of >7 m or most intersection lies in outside valid area), while the NLS algorithm performs better than Intersection Density; it has the hit ratio (55 – 70%) for the real-time data. The use of KF can improve the hit ratio of 81.15 %.

According to the experiment, we assume that our algorithm still has a low-accuracy for the real-time data case; it is because of the RSSI instability. In the next work, we will try to overcome this problem by using the Unscented Kalman Filter (UKF) to pre-process the RSSI. We can put the UKF algorithm after the localization result to get better results.

References

[1] V. Vinhas, "Business Intelligence Based on A Wi-Fi Real Time Positioning Engine A Practical Application In A Major Retail Company," Proc. of the 11th Int. Conf. on Enterprise Information Systems, 2009.

[2] Y. Yuan, et al., "Crowd Monitoring Using Mobile Phones," Proc. of the 6th Int. Conf. on Intelligent Human-Machine System and Cybernetics, 2014.

[3] A.B.M. Musa and J. Eriksson, "Tracking Unmodified Smartphones using Wi-Fi Monitors," Proc. of the 10th ACM Conf. on Embedded Network Sensor Systems, pp. 281-294, November 2012.

[4] M. Cunche, "I know your MAC address: targeted tracking of individual using Wi-Fi," J. of Computer Virology and Hacking Techniques, Springer, 2013.

[5] S. Fuada, et al., "Your MAC Address can be detected easily when connected to the Wi-Fi," Unpublished.

[6] J. Weppner, et al., "Monitoring Crowd Condition in Public Spaces by Tracking Mobile Consumer Devices with WiFi Interface," Proc. of the UbiComp, pp. 1363-1371, 2016.

[7] Y. Fukuzaki, et al., "A Pedestrian Flow Analysis System using Wi-Fi Packet Sensors to a Real Environment," Proc. of the UbiComp, pp. 721-730, 2014.

[8] N.E. Gemayel, et al., "A Weighted Least Squares Algorithm for Passive Localization in Multipath Scenarios," Proc. of VTC2015-Fall, 2015.

[9] C.Yang and H. Shao, "WiFi-Based Indoor Positioning," IEEE Communication Magazine, Vol. 53(3), pp. 150-157, 2015.

[10] R. Ma, et al., "An Improved WiFi Indoor Positioning Algorithm by Weighted Fusion," Sensors, pp. 21824-21843, Vol. 15(9), 2015.

[11] N. Hernandez, et al., "WiFi-based indoor localization and tracking of a moving device," Proc.of UPINLBS, 2014.

[12] L. Khalil and P. Jung, "Scaled Unscented Kalman Filter for RSSI-Based Indoor Positioning and Tracking," Proc. of NGMAST, 2015.

[13] M. Shchekotov, "Indoor localization methods based on Wi-Fi lateration and signal strength data collection," Proc. of the 17th of fruct association, pp. 187-191, 2015.

[14] G. Pipelidis, et al., "HuMan: Human Movement Analytics via WiFi Probes," Proc. of PerCom Workshops, 2019.

[15] D.N. Fernandez, et al., "Implementation of a WiFi-based indoor localization system on a mobile device for a university area," Proc. of Intercon, 2019.

Priority-based Routing Framework for Image Transmission in Visual Sensor Networks: Experimental Analysis

Emad Felemban[1]
College of Computing and Information Systems
Umm al Qura University
Makkah, Saudi Arabia

Atif Naseer[2]
Science and Technology Unit
Umm al Qura University
Makkah, Saudi Arabia

Adil Amjad[3]
Science and Technology Unit
Umm al Qura University
Makkah, Saudi Arabia

***Abstract*—A Visual Sensor Network (VSN) is a specialized Wireless Sensor Network (WSN) equipped with cameras. Its primary function is to capture images, videos and send them to power rich sink nodes for processing. As image data is much larger than scalar data sensed by a typical WSN, applications of VSN require much bigger amount of data to be transferred to the sink. Due to constraints of WSN such as low energy, limited CPU power and scarce memory, transmission of large amount of data becomes challenging. On the other hand, some VSN applications require critical image features sooner than the entire image to take action. In this paper, we provide the details of experiments done using our proposed Priority-based Routing Framework for Image Transmission (PRoFIT). PRoFIT is designed to deliver critical image features at high priority to the sink node for early processing. Peak signal-to-noise ratio (PSNR) analyses show that PRoFIT improves VSN application response time as compared to priority-less routing. This paper also contains the design of our VSN testbed. Multiple indoor and outdoor experiments were performed to validate the framework. This framework also improves the energy efficiency of the network. The results show that the PRoFIT is 40% efficient in terms of energy consumption.**

***Keywords*—Priority-based routing; visual sensor networks; testbed; framework**

I. Introduction

Visual Sensor Networks (VSNs) are wireless sensor networks (WSN) equipped with image sensors or cameras. Typical WSNs sense and transmit environment factors, whereas VSNs sense and transmit images that are considerably larger than typical WSN data. VSN applications, such as periphery monitoring, intruder detection, surveillance, etc. require very large amounts of image data to be transmitted from camera nodes to sink. In most cases, the memory size required to store and send is 16-bits per reading. On the other hand, a VSN node, equipped with a camera generates vector data. For instance, a raw Red-Green-Blue (RGB) image of 128 x 128 pixels with 24-bits per pixel (8 bits per color) will be of 128 x 128 x 24 = 393216 bits (approximately 48 kilobytes). These are magnitudes larger than traditional sensor data.

Some VSN architects use image compression techniques such as Discrete Cosine Transforms [1] or Discrete Wavelet Transforms [2], yet these techniques do not reduce data size significantly. Other methodologies such as progressive image transmission [3] and bit plane coding [4] transmit image in multiple layers. These methodologies partition the image into layers such that the primary layer contains the most significant features of the image, whereas successive layers contain details that when integrated with the primary layer, reconstruct the complete image. Each layer is of a fraction in size of the complete image. If packets of the primary layer are transmitted at a priority through a VSN, they will reach the sink node much sooner than the complete image. Some VSN applications may improve response time by processing the primary layer(s) first instead of waiting for the complete image to be received. This motivated us to develop a priority based routing framework for VSNs called PRoFIT. The concepts, design and benefits of PRoFIT was demonstrated using simulations on Cooja Simulator in our previous work [5]. We proved that PRoFIT delivers critical information required for multimedia applications much sooner than priority-less routing frameworks. Then, we focused on end-to-end delays and packet delivery ratios at different deadlines using PRoFIT [6].

In this paper, we designed a real wireless network testbed to analyze the performance of PRoFIT algorithm[5] in indoor and outdoor environment. This testbed was designed to capture images using a controllable PTZ camera. The captured image can be sent over multihop network using ZigBee. In addition to using PRoFIT, researchers can evaluate their custom protocols, frameworks, algorithms and VSN applications on testbed. In this paper, we present Peak signal-to-noise ratio (PSNR) analyses to show that PRoFIT improves VSN application response time as compared to priority-less routing on real testbeds. We also measured the energy efficiency of the network and show that the PRoFIT is 40% efficient in terms of energy consumption.

The rest of the paper is sectioned as follows. Related work is provided in Section 2. Priority Based Routing Framework for VSN is discussed in Section 3. The PRoFIT implementation on real testbed is explained thoroughly in Section 4. We evaluated PRoFIT on real testbed. This served as an example that researchers can use to evaluate their own VSN applications and protocols. The details of the experiments and results are discussed in Section 5. Finally, the paper is concluded in Section 6.

II. Related Work

Mostly, authors have focused on different aspects of VSN performance, such as network lifetime. Zhou et al. [7] propose fast visual background extraction algorithm to detect motion and then use prioritized transmission to balance between delay and energy consumption. Han et al. [8] have surveyed routing protocols that balance out the trade-off between network lifetime and QoS requirements. They provide a baseline for "green routing protocols". Similar to ours, Bondi et al. [9] provided a framework that is implemented on real VSN nodes. They advocate on comparing analyze-then-compress (distributed approach) verses compress-then-analyze (centralized approach).

Eriksson et al. [10] also demonstrate that distributed optimization can improve system response time for VSNs. Bhosale et al. [11] proposed prioritized scheduling technique at internode level to maximize the lifetime of network. Spachos et al. [12][13] proposed a dynamic routing protocol called Content Relevance Opportunistic Routing (CROR) for improving overall network performance by focusing on Quality of Service parameters such as energy efficiency and packet delay.

WSN research teams have been using testbed to experiment and validate their suggested architectures and protocols. They have designed such testbed with different architectures, hardware selection, deployment, integration, remote management and various capabilities. The survey in [14] has aggregated and classified many WSN testbed that have been implemented until the time for publication in 2014. The classification were based on i) research focus, ii) architectural aspect and requirements and iii) experimentation nature on testbeds.

The latest survey that included testbed literature review was published in 2008 by Akylidiz et al. in [15]. The authors surveyed different testbed architecture, design space, hardware and finally, listed out all features provided by surveyed testbeds for VSN. In [16], the authors implemented a testbed for mobile robotics and heterogeneous wireless sensor nodes with different sensing capabilities including cameras. The testbed is indoor with static and mobile sensor nodes that are based on classical sensor node platforms plus mobile robotics. The testbed has been used to verify and experiment with many projects including tracking and communication testing.

We noticed that VSN testbed are mainly used to verify methodology and research outcome for computer vision related research such as coverage, encoding and compression techniques. Some of them are also used to experiment with other research problems such as wireless communication, networking and power efficiency. We noticed also that most testbed are indoor and some of them are miniature in size. In [17] the authors developed a compression technique to speed up the wireless transmission of data. By sending a fast compressed and specially encoded scan of the image, the authors claim that the network can deliver the initial version of the image very quickly to take fast decisions. The testbed used composed of two imote-2 nodes and X-scale sink node. In [18] the authors used a network of camera nodes in a miniature testbed to test their algorithm for camera coverage. In [19], the authors used a networked camera nodes testbed to maximize the field of view of all cameras in the network.

III. Priority-based Routing Framework for VSN

Typically the network model used in VSN consists of camera nodes, intermediate nodes and sink nodes. Fig. 1 shows a sample surveillance network. We assume typical VSNs to be of this type. That is, camera nodes will be at a periphery, whereas intermediate nodes will relay image information from camera nodes to the sink.

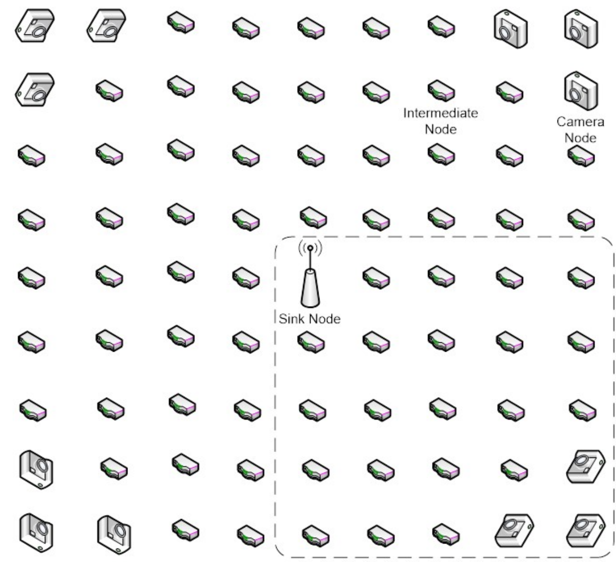


Fig. 1. VSN Network Model

Functionality of our PRoFIT framework is distributed into network layer and medium access control layer of any protocol stack as depicted in Fig. 2 .Additionally, a thin Application Interface Layer (AIL) encapsulates the details of network layer and medium access control layer. The subsections below are provided the functional details of these layers.

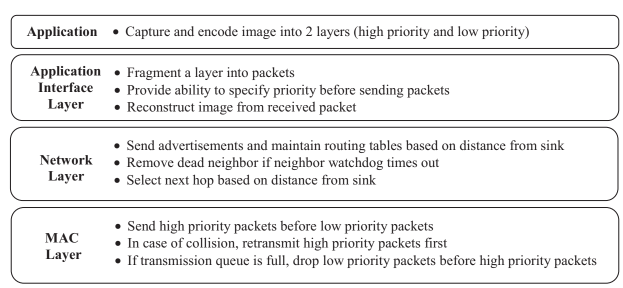


Fig. 2. VSN application and PRoFIT layers

A. Application Interface Layer

The AIL (Application Interface Layer) is the application layer component of priority-based routing framework. It is a very thin layer that provides VSN application with a set of primitives that can be used for fragmenting image data into packets, sending them, receiving them and assembling them to re-generate image data. The AIL hides the implementation details of the entire framework. The VSN application passes image data along with its priority to the routing framework using the AIL. Based on its configuration, AIL of the sending node fragments the image data into packets of size that network layer can send. AIL also inserts image number and packet fragment number into the packet. This information is used by the AIL of sink node to join the fragments to construct image data sent.

B. Network Layer

The network layer component of PRoFIT works in two phases explained below.

1) Network Configuration Phase

When the VSN is deployed and brought up, the VSN nodes send advertisements to their neighbors declaring identities and their number of hops from sink. These advertisements are sent periodically. Initially all nodes are configured as being infinitely away from sink node. When sink node advertises, it declare its number of hops from sink as 0. The nodes receiving this advertisement add the respective sink node to their routing tables and mark their number of hops from sink as one hop. Now when such a node sends out its own advertisement, it declares its number of hops from the sink instead of infinity. The nodes at multiple hops from sink update their routing table with sink address along with the addresses of their neighbor as next hop address from who they received the advertisement. When a node receives advertisement of a sink from more than one neighbor, it keeps only the neighbor with lesser hops to the sink in its routing table. After a number of cycles of advertising, depending on the number of VSN nodes, the network is established. Each node knows the number of hops to the sink as well as the next hop towards the sink. As the advertisements are sent out periodically, removal and addition of nodes to the network is possible dynamically. Moreover, for maintenance of routing tables, each node keeps track of live neighbors using a watchdog timer associated with each neighbor.

2) Network Operation Phase

Once the network has been established, our routing framework is ready to transport image data from camera nodes to sink nodes. When the VSN application has image data to send, it uses primitives provided by the AIL from previous section. The network layer selects the next hop towards the sink that is selected by the camera node from its routing table. If the sink address as specified by the camera node is not in the routing table, the packet is dropped. A neighbor's entry keep-alive watchdog is reset whenever a packet is received from that neighbor. If a packet is not received from a neighbor within a threshold, the neighbor's entry is deleted from the routing table. In this way, routing tables are maintained during data transmission phase.

C. Medium Access and Control Layer

At the MAC layer, the routing framework works at two levels. The first is the intra-node level where the routing framework makes sure that high priority packets are forwarded before low priority packets. The second level is the inter-node level where the routing framework makes sure that when two neighbors contest for transmission medium, the neighbor with high priority packet gets a chance to transmit its packet before the neighbor with low priority packet. The following sub-sections explain these two levels.

1) Queue Insertion

When a packet arrives at MAC layer for transmission, it is sent instantaneously if the MAC layer is not already receiving or sending a packet. If the MAC layer is busy, the packet is placed in a queue where it waits for its turn. Our priority-based routing framework makes use of this queue. When a packet with high priority arrives, it is placed at the head of the queue so that it is sent in the next go. If a packet of low priority arrives, it is placed at the tail of the queue. As the MAC layer always selects packets from head of the queue for transmission, it is made sure that at intra-node level a packet with higher priority is transmitted first.

2) Differentiated Back-off Window

When two nodes find the medium available and transmit at the same time, a collision occurs. In regular CSMA/CD, both nodes back off for a randomly selected time slot from a pseudo-fixed-size window. If they collide again, the window size is increased exponentially to a certain size. The priority-based routing framework maintains different windows for the different priorities. When a collision occurs, the MAC layer checks the priority of packet that collided and determines back-off times from different windows. For high priority packet, the window is smaller than for a low priority packet. This way, if the node with high priority packet gets a chance to transmit its packet within a smaller window than a node with a low priority packet. This makes sure that at the inter-node level, high priority packets transmit sooner than low priority packets.

IV. PRoFIT Implementation

Fig. 3 shows the routing framework data flow. The topmost block represents the VSN application and its usage of AIL. The middle block represents packet en-queueing into MAC layer transmission queue. The bottom block signifies the transmission of packet and calculation of contention window in case of collision. To quantify the usefulness of the routing framework, a real testbed has been deployed and experimentation were run. PRoFIT is implemented on three types of nodes. *(1)* The first type of nodes called camera nodes are equipped with image sensor. *(2)* The second type are intermediate nodes that may or may not take part in sensing but are used as relay nodes. *(3)* The third type are sink nodes that process received images and help the application in making decisions.

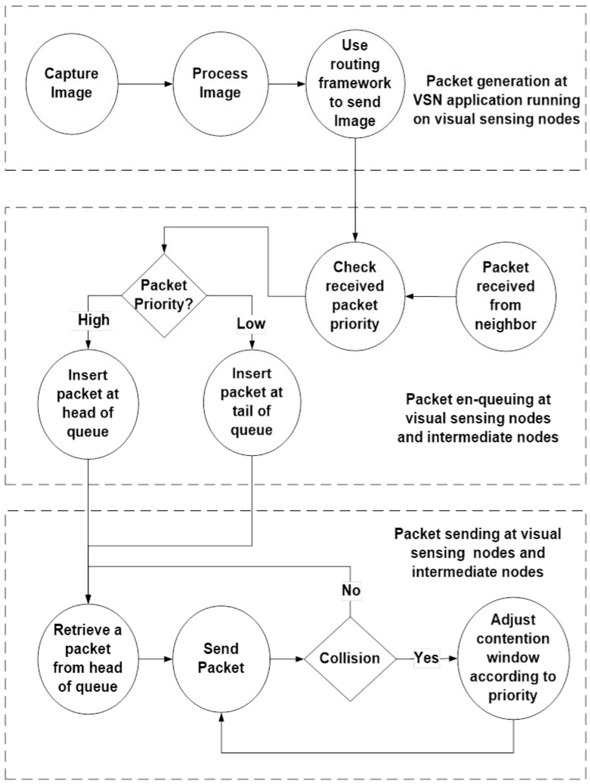


Fig. 3. PRoFIT Framework Data Flow

A. Camera Node

Fig. 4 show the processes running on the camera nodes. The camera node captures the image and transmits it in four steps.

Step 1 (Encoding): For a captured grayscale image, 1 byte (8 bits) is used to represent the gray level of one pixel. PRoFIT encodes pairs of pixels of the captured image and stores them into two layers. In the current version of PRoFIT, bit plane encoding [4] is used to encode the image into two layers; primary layer and secondary layer. The primary layer consists of bit planes 8 to 5 of a pixel of the captured image while the secondary layer consists of bit planes 4 to 1 of the same pixel. This mechanism is shown in Fig. 5. For example, the first and second pixels of the captured image are partitioned into four sets of four bits each. The four most significant bits (MSBs) of the first pixel of captured image are stored at the four MSBs of first pixel of primary layer. The four MSBs of the second pixel of captured image are stored at the four least significant bits (LSBs) of first pixel of primary layer. The four LSBs of the first pixel of the captured image are stored at the four MSBs of the first pixel of the secondary layer. The four LSBs of the second pixel of the captured image are stored at the four LSBs of the secondary layer. Similarly, the third and fourth pixel of the captured image becomes second pixel of each layer, and so on. In this way, each encoded pixel of primary layer contains the MSBs of two pixels of captured image, whereas each encoded pixel of the secondary layer contains LSBs of the the same two pixels of captured image. Note that the size of both encoded layers combined is same as the original image.

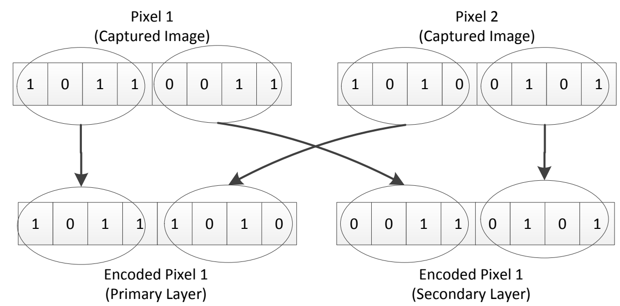


Fig. 5. Encoding Captured Image Pixels into Two Layers

Step 2 (Fragmentation): Packet size of VSN physical layer may be limited. As image data is much larger, it must be fragmented into packets of size that the physical layer can transit. Maximum packet payload size is 72 bytes in a typical sensor network that uses Zigbee for transmission. Hence, the current version of PRoFIT has been designed with IEEE 802.3 (Ethernet) for transmission with a packet size of 72 bytes. The physical layer can be replaced in the future versions. PRoFIT packet contains three fields; flags, sequence number and payload. Sequence number field is required in order to defragment at the sink. Sequence number field of two bytes can accommodate QCIF (176×120) to 4 MP (2240×1680) image resolution formats. A QCIF image would be fragmented into 307 fragments, where as 4MP image will be fragmented into $54,540$ fragments. One byte of the packet represents various flags such as priority, new image, etc. The remaining 69 bytes are used as payload to send encoded pixels.

Step 3 (En-queuing Packets): PRoFIT contains two queues; A high-priority packet's queue called Queue-1 and a low priority packet's queue called Queue-2. As the primary and secondary layers of the image are being packetized (Step 2), each packet is inserted at respective queue's tail. The packets of primary layer are inserted into Queue-1 and packets of secondary layer are inserted into Queue-2.

Step 4 (Dispatching Packets): The packet dispatcher runs in a loop as a child thread of the main program. Fig.4 depicts the mechanism of selecting packets from the queues and transmitting them. If there is a packet in Queue-1, then the packet at the head of Queue-1 is extracted for transmission. If Queue-1 is empty and there is a packet in Queue-2, it will be selected for transmission. If both queues are empty, the dispatcher waits for a packet to arrive in one of the queues. Using the above four steps, high priority packets are always delivered before low priority packets.

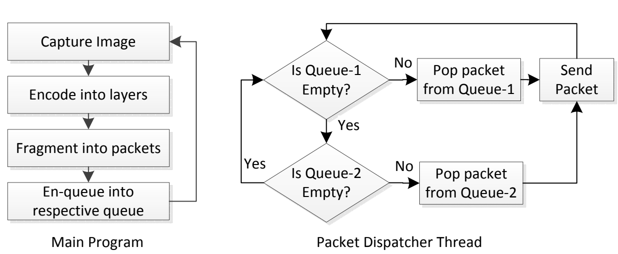


Fig. 4. Processes running on camera node

B. Intermediate Node

In our network, the intermediate nodes are only responsible for routing packets from camera nodes to sink node. They do not take part in sensing or sharing processing load of the camera nodes or sink nodes. When the network is deployed, the intermediate nodes create routing tables that are necessary to take routing decisions when packets are received. To achieve their primary task of routing image data from visual sensing nodes to sink nodes, the routing tables in intermediate nodes are updated throughout the lifetime of the network as some nodes may die due to depleted power or other environmental factors, while other nodes may be added to the network when required. PRoFIT makes sure that intermediate nodes forward high priority packets faster than low priority packets. In this way, PRoFIT facilitates sink nodes to reconstruct an image using high order bit planes much sooner than when the entire image data is received at sink. As required, the sink node can add lower order bit planes to the first pass to construct a more detailed image.

C. Sink Node

The sink node uses two steps to reconstruct image.

Step 1 (Defragment Packets): Based on a preset image resolution, the sink pre-allocates memory to store incoming encoded pixels into the primary and secondary layers. The sink node uses the sequence number field to defragment. Packet with sequence number $1, 3, 5$ and so on are part of primary layer, whereas packets with sequence number $2, 4, 6$ and so on are part of secondary layer. Each packet contains 69 encoded pixels. As a packet is received, the sequence number is extracted to find the location of the encoded pixels in the respective layer. Encoded pixels from the packet with sequence number 1 are stored in place of the first 69 pixels of the primary layer. Encoded pixels from the packet with sequence number 3 are stored in place of the second 69 pixels of the primary layer, and so on. Encoded pixels from the packet with sequence number 2 are stored in place of the first 69 pixels of the secondary layer. Encoded pixels from the packet with sequence number 4 are stored in place of the second 69 pixels of the primary layer, and so on.

Step 2 (Decode Packets): On a preset periodic deadline, the sink node decodes the primary and secondary layers by splitting each encoded pixel and re-assembling the bit planes. The first pixel of the primary and secondary layer are split into four sets of four pixels each. The 4 MSBs of first pixel of primary layer are concatenated with 4 MSBs of the first pixel of secondary layer to reconstruct first pixel of sent image. The 4 LSBs of first pixel of primary layer are concatenated with 4 LSBs of the first pixel of secondary layer to reconstruct second pixel of the sent image. The 4 MSBs of second pixel of primary layer are concatenated with 4 MSBs of the second pixel of secondary layer to reconstruct third pixel of sent image. The 4 LSBs of second pixel of primary layer are concatenated with 4 LSBs of the second pixel of secondary layer to reconstruct fourth pixel of the sent image, and so on. At any given time, more packets of the primary layer than secondary layer would have been received at sink. Therefore, almost half way through image transmission, critical image features are available at the sink node for processing.

V. Experimentations and Results

The motivation to implement the PRoFIT algorithm over the testbed is to obtain the desired image in the shortest time with minimal energy consumption. The quality of the received image was measured by the PSNR. We performed multiple sets of experiments to measure the quality of the received image as well as energy efficiency of PRoFIT. PRoFIT was evaluated on the real testbed in indoor and outdoor settings. The following subsections contain the details of the experiments.

A. Testbed Design

To implement the PRoFIT on real testbed, we design a testbed consisting of VSN nodes. The following requirements have been compiled for an energy efficient VSN node.

Camera: VSN node should have a camera or image sensor to capture visual information. The camera should be able to capture images in different lighting conditions. Roseek Cheetah1 Camera System is used for camera node [20]. The view of VSN node's camera should be able to pan, tilt and zoom (PTZ) in order to maximize coverage. Therefore, the camera was installed on a PTZ modules.

Battery: The VSN node should consume power from a battery unit that is able to be charged using energy harvesting modules. Two deep-cycle gel batteries without of 12 V at 100 Ah were used to provide 48 hours of power backup. Unlike traditional WSNs, VSN node requires much higher amount of energy. The energy harvesting unit of VSN node should harvest enough energy to power the VSN node as well as charge its battery to run during night time. SUN-5 Inverter and Charge Controller with PWM of 1 KVA and 12 V was used with 250 W PV solar panel. The inverter provided power to both CPU and camera.

Communication Unit: The testbed VSN node would require at least two communication interfaces; First one based on Zigbee for VSN data transmission at low power and the second one for controlling and debugging VSN node. The latter can use any other wired or wireless technology. It will only be used to setup the VSN, configure it and debug the VSN node operations. In normal VSN working the latter will be disabled to conserve energy. ZigBee - XBP24BZ7SIT-004 was interfaced with CPU using ZigBee USB adapter board and ZigBee 5 dBi RP-SMA antenna was used for image transmission communication unit. PRoFIT was interfaced with Open Source ZigBee drivers to send/receive image data to/from ZigBee module. USB Wi-Fi Adapter with 4-inch antenna was interfaced with CPU as control and debug interface.

Operating System and Open Source Software: Roseek Cheetah1 series comes with Linux Ubuntu distribution, OpenCV and Roseek's own image processing SDK. Image acquisition was carried out using OpenCV and PRoFIT was used to partition the image into coarse and fine image detail layers. The coarse layer containing critical image data was transmitted at high priority using PRoFIT where as fine image information was transmitted at low priority using PRoFIT.

B. Indoor Experiments

To set up an indoor testbed for a multi-hop network, three nodes were used for experimentation. PRoFIT was ported for all three nodes types; namely, camera node, intermediate node and sink node. The camera node captured images on demand and used PRoFIT to prioritize the coarse image information. Then it sent the packets to intermediate node. The intermediate node also used PRoFIT to relay incoming packets. The sink node received packets from the intermediate node and used PRoFIT to defragment them and decode them. Fig. 6, shows the testbed of multi-hop indoor experimentation. Two standard images were selected for experimentation. These images were obtained from The Standford Center for Image Systems Engineering (SCIEN). The images were baboon.tif with 256 x 256 resolution and lena.tif with 512 x 512 resolution. Each hop was 30 meters away from the other. Multiple experiments were run to transmit these images from the source node to the sink node

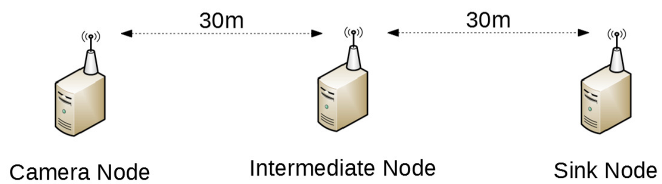


Fig. 6. Multi-hop Indoor Testbed

1) Time Evaluation: To evaluate the received images with respect to time, multiple sets of experiments were performed with single-hop and multi-hop setups. Timestamps of received image packets were used. Once the image PSNR reached 0.9, the time was noted. Such high PSNR is good for applications for decision making. Below is the comparison of different experiments with and without PRoFIT algorithms.

In the first set of experiments, the reception time was evaluated over single hop; firstly, by using PRoFIT and then without using PRoFIT. The PSNR of both received images were analysed. Fig. 7 shows the comparison of PSNR of the image *baboon.tif (256 x 256)*. The graph shows a clear difference between the received images PSNR. The image received where PRoFIT was enabled had 0.9 PSNR at 24,000 microseconds, whereas the same image received without PRoFIT had 0.9 PSNR at 49,000 microseconds. Similar results are shown in Fig. 8 in which the PSNR of image*lena.tif (512 x 512)* was evaluated. The graph shows that the PSNR of the image received with PRoFIT reached to PSNR 0.9 at 40,000 microseconds. However, without PRoFIT this image reached to 0.9 PSNR at 49,000 microseconds. Hence, these experiments prove that the image received using PRoFIT takes considerably less time than normal transmissions without the PRoFIT.

Fig. 9 shows the received images when PSNR reached 0.9. Fig. 9 (a) and Fig. 9 (c) are the images received with the PRoFIT, whereas Fig. 9 (b) and Fig. 9 (d) show images received without PRoFIT. The images received using PRoFIT at PSNR 0.9 are almost complete and can be used in any decision-making, whereas, the images received without PRoFIT at same PSNR are almost half-complete. PRoFIT impacts approximately 50% in image reception time.

The next set of indoor experiments were performed to evaluate the received images PSNR over multi-hop. The same

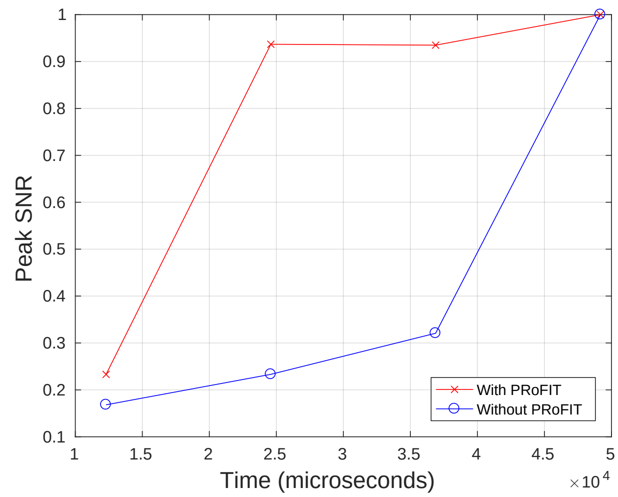


Fig. 7. 256x256 With PRoFIT vs without PRoFIT single-hop

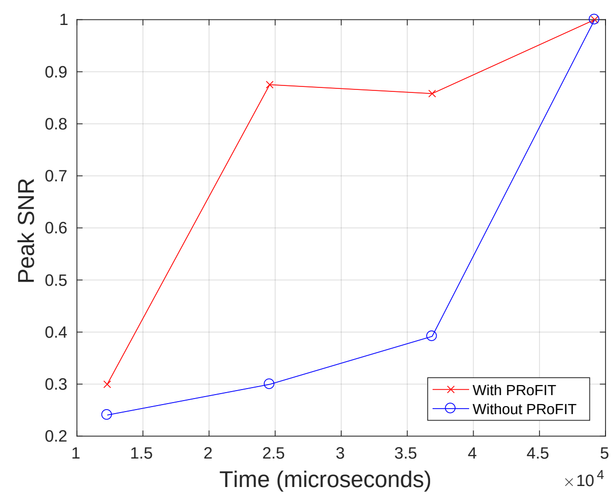


Fig. 8. 512x512 With PRoFIT vs Without PRoFIT (single-hop)

images were transmitted with PRoFIT and without PRoFIT. Fig.10 shows the PSNR evaluation of 256 x 256 image using multi-hop transmission. The image transmitted with PRoFIT reached to 0.9 PSNR at 25,000 microseconds, whereas the same image transmitted without PRoFIT reached to 0.9 PSNR at 49,000 microseconds.

Fig. 11 shows the PSNR value of the 512 x 512 image. The PSNR of the received image with PRoFIT was higher right from the start compared to the image without PRoFIT. The PSNR reached to 0.9 at 40,000 microseconds with PRoFIT, whereas without PRoFIT the PSNR reached at 49,000 microseconds. Both multi-hop graphs show that, in multi-hop transmission, the PSNR of received images with PRoFIT reached 0.9 PSNR much sooner compared to without PRoFIT.

Fig. 12 shows the received images when PSNR reached 0.9. Fig. 12 (a) and Fig. 12 (c) are the images received with PRoFIT using multi-hop transmission. Fig. 12 (b) and Fig. 12 (d) shows the images received without PRoFIT using multi-

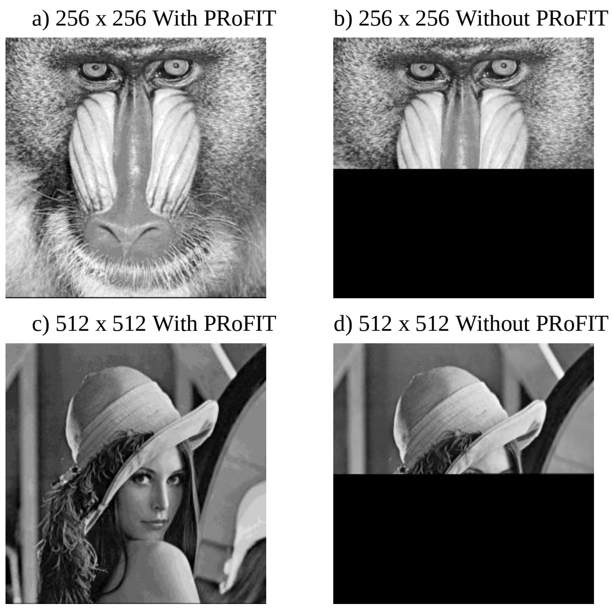


Fig. 9. With PRoFIT vs Without PRoFIT (single-hop)

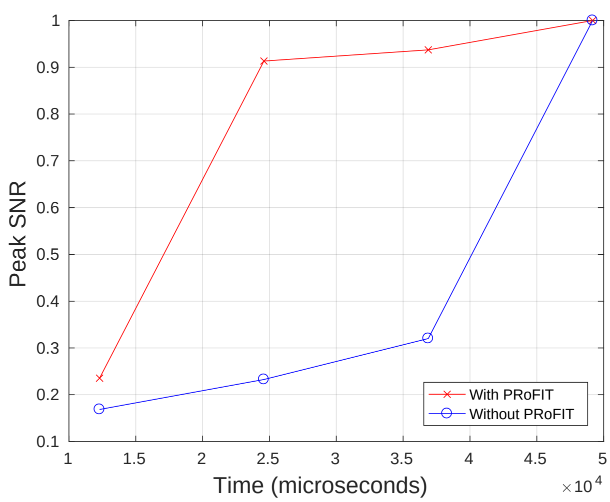


Fig. 10. 256x256 With PRoFIT vs without PRoFIT Multi-hop

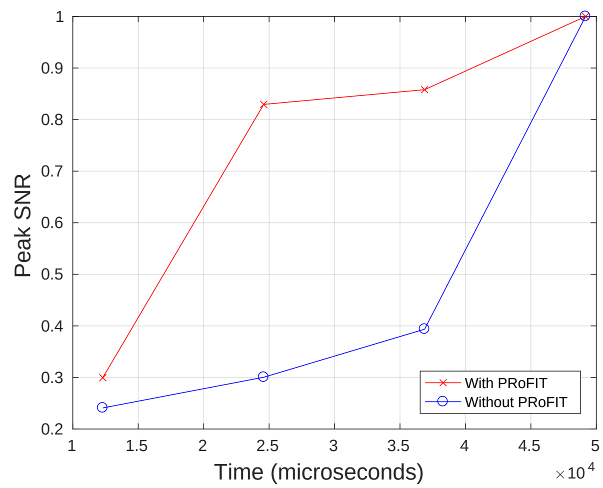


Fig. 11. 512x512 With PRoFIT vs without PRoFIT Multi-hop

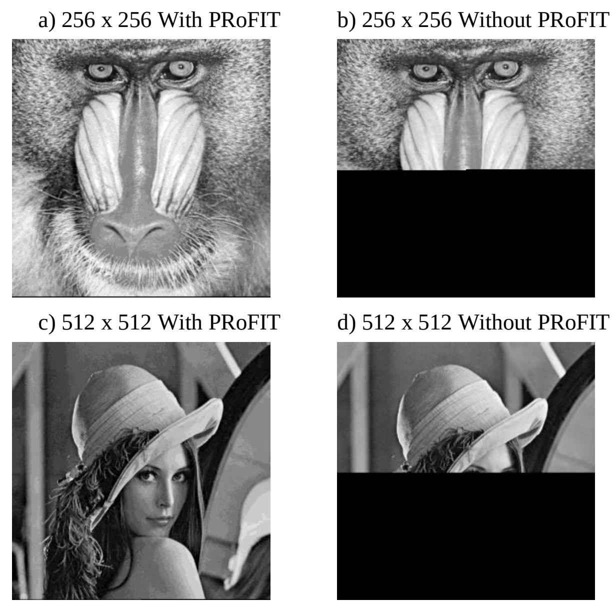


Fig. 12. WithPRoFIT vs WithPRoFIT Multi-hop

hop transmission. The images received with PSNR 0.9 through multi-hop transmission using PRoFIT are almost complete and can be used in any decision-making. The images received without PRoFIT with the same PSNR are almost half complete. Therefore, PRoFIT in multi-hop transmission also impacts 50% in image receiving time.

In terms of time evaluation, it is seen through the indoor experiments that image received with PRoFIT enabled reached 0.9 PSNR much sooner than the image received without PRoFIT enabled. Hence, PRoFIT impacts greatly on time in both single-hop and multi-hop indoor transmissions.

2) Energy Evaluation: The evaluation of energy consumed is very crucial in VSNs. We evaluated energy consumption of images received in a multi-hop indoor environment. The results were compared using with PRoFIT and without PRoFIT transmissions. Fig. 13 shows the relationship between number of packets received against accumulated PSNR. It clearly shows that the PSNR increases as the number of packets are received, but, the significant difference is when PRoFIT is enabled. As in Fig. 13, it is clearly noted that with PRoFIT, the PSNR value of 0.9 was achieved after receiving 500 packets (45% of the total packets) when PRoFIT was enabled, whereas in the test with without PRoFIT, the same PSNR was achieved after receiving 1000 packets (91% of the total packets). Hence, we can easily relate the energy consumed in both transmissions. The decision can be easily taken after 0.9 PSNR of the received image, so 45% of energy can be saved using PRoFIT as compared to transmission without the PRoFIT.

Fig. 14 shows the energy consumption of the indoor image *lena.tif (512 x 512)* transmission with and without PRoFIT. The graph shows the relationship between the number of packets

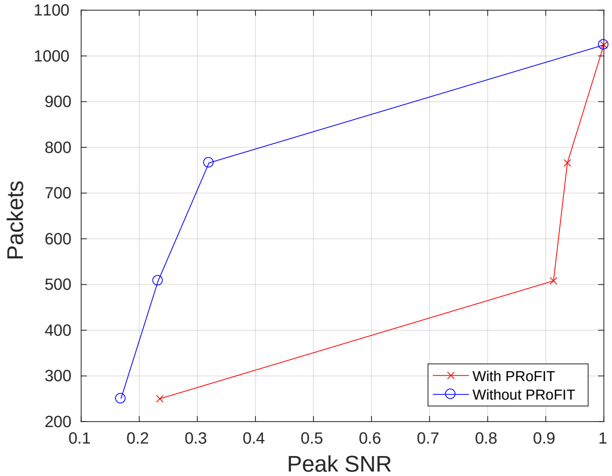


Fig. 13. Energy Consumption of image 256x256 Multi-hop Indoor Transmission

received and accumulated PSNR of the image. The image with PRoFIT reached the PSNR value of 0.9 when 2000 packets were received (48% of total packets), whhereas the image without PRoFIT reached the PSNR value of 0.9 when 3700 packets were received (83% of total packets). Therefore, the energy consumed with the PRoFIT was 35% less than the energy consumed without the PRoFIT in indoor environment.

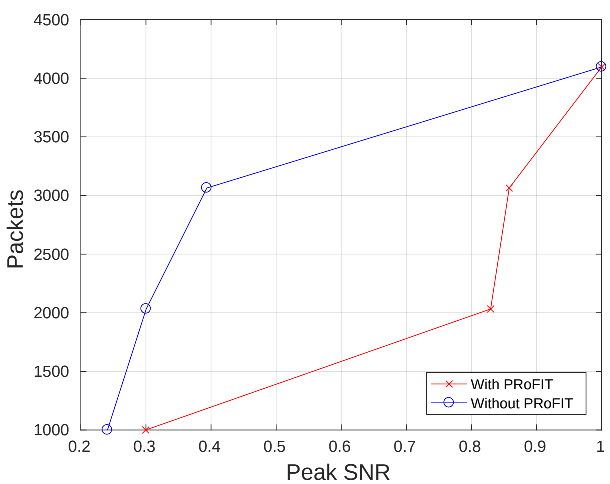


Fig. 14. Energy Consumption of image 512x512 Multi-hop Indoor Transmission

C. Outdoor Experiments

To set up an outdoor testbed for a multi-hop network, three nodes were used for the experiment. The testbed was deployed in one of the open air car parks at the Umm al Qura University, Saudi Arabia. Same PRoFIT implementation from indoor experiments was used. The cameras were attached with each node. Fig. 15 shows the testbed of the multi-hop outdoor experiment. Multiple experiments were performed on the outdoor testbed. Baboon.tif image was used for outdoor experiments. Every hop was 30 meters away from the other. Multiple experiments were carried out to transmit this image from the camera node to the sink node.

Fig. 15. Multi-hop Outdoor Testbed

1) Time Evaluation: To evaluate the received image with respect to time, multiple sets of experiments were performed with single-hop and multi-hop set-up in the outdoor environment. The PSNR value was calculated for the received image. Once the image PSNR reached 0.9, it provided a preview that could be used for decision-making. Below are the comparison between the different experiments with and without the implementation of the PRoFIT algorithms.

In the outdoor environment, the first set of experiments were performed to evaluate time over single-hop with and without the PRoFIT. Fig. 16 shows that PSNR values reached 0.9 at 24,000 microseconds with PRoFIT enabled transmission. However, the PSNR reached 0.9 at 47,000 microseconds without the PRoFIT enabled transmission.

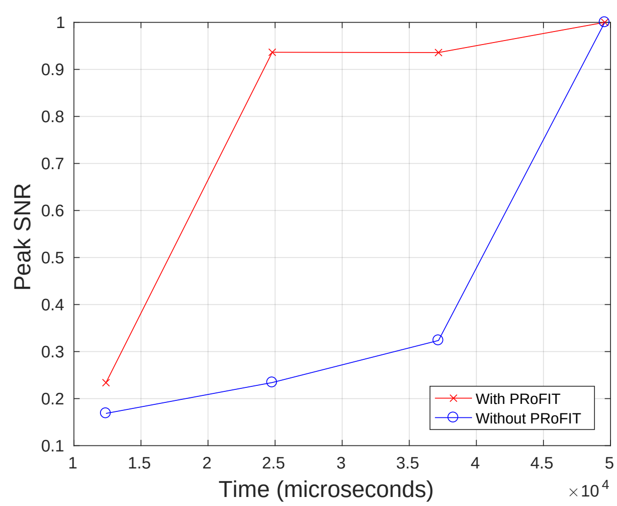


Fig. 16. 256 x 256 With PRoFIT vs Without PRoFIT Single-hop Outdoor

Fig. 17 shows the received images when PSNR reached 0.9. Fig. 17 (a) is the image received with single-hop with PRoFIT, whereas Fig. 17 (b) shows the image received without PRoFIT. The image received with PRoFIT reached the sink node almost completely at PSNR 0.9. However, the image received without the PRoFIT algorithm reached half only. Therefore, the image transmitted using PRoFIT saves 50% of image transmission time.

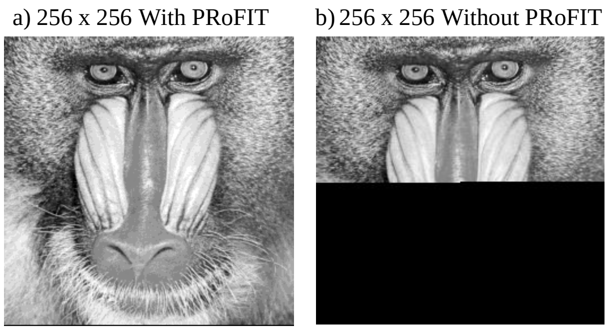


Fig. 17. 256 x 256 With PRoFIT vs Without PRoFIT Single-hop Outdoor

The second set of experiments were performed to analyze the image transmission in an outdoor environment over multi-hop network. The image, *baboon.tif (256 x 256)*, was transmitted using multi-hop with and without PRoFIT. Fig. 18 shows the comparison between the two received images' PSNR. Over multi-hop with PRoFIT, the value of PSNR reached 0.9 in 52,000 microsecond. Over multi-hop without PRoFIT, the PSNR reached to 0.9 in 72,000 microseconds.

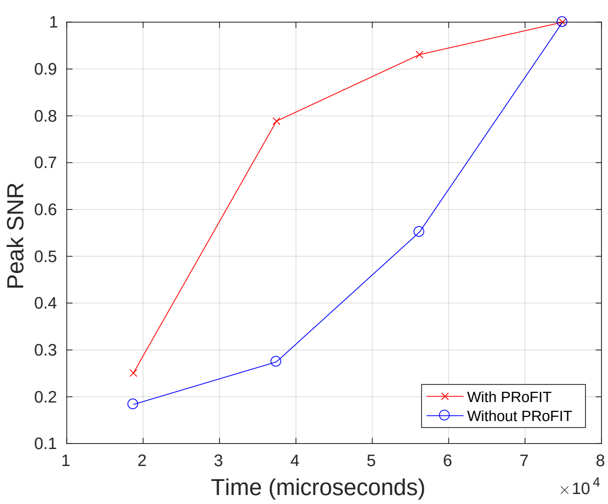


Fig. 18. 256 x 256 With PRoFIT vs Without PRoFIT Multi-hop Outdoor

Fig. 19 shows the received images when PSNR reached 0.9. Fig. 19 (a) is the image received with multi-hop with PRoFIT, while Fig. 19 (b) shows the image received without PRoFIT. The image received using PRoFIT reached almost 90% to sink node at PSNR 0.9, while the image received without PRoFIT reached only 50% of the complete image. The image transmitted using PRoFIT saved almost 40% of the image transmission time.

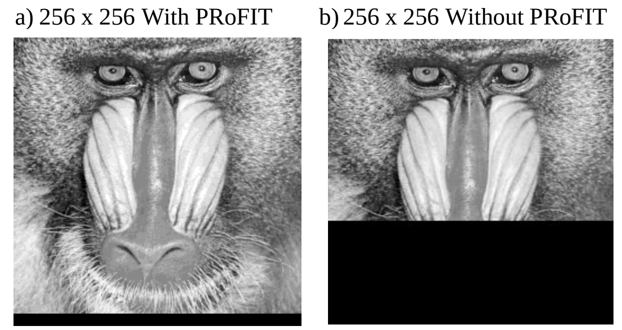


Fig. 19. 256 x 256 With PRoFIT vs Without PRoFIT Multi-hop Outdoor

2) Energy Evaluation: This section briefly discusses the evaluation of energy in outdoor image transmission. Fig. 20 shows the relationship between number of packets received against accumulated PSNR. It clearly shows that PSNR increases with the number of packets received, but, the significant difference is when the transmission is done using PRoFIT. In Fig. 20, it is clearly noted that with PRoFIT the PSNR value of 0.9 is achieved after receiving 500 packets (50% of the total packets). Without PRoFIT, the same PSNR was achieved after receiving 900 packets (90% of the total packets). Hence, we can easily relate the energy consumed in both transmissions. One can easily identify the image and make any decisions once the image PSNR reaches 0.9, so we can save 40% of energy using PRoFIT as compared to transmission without PRoFIT.

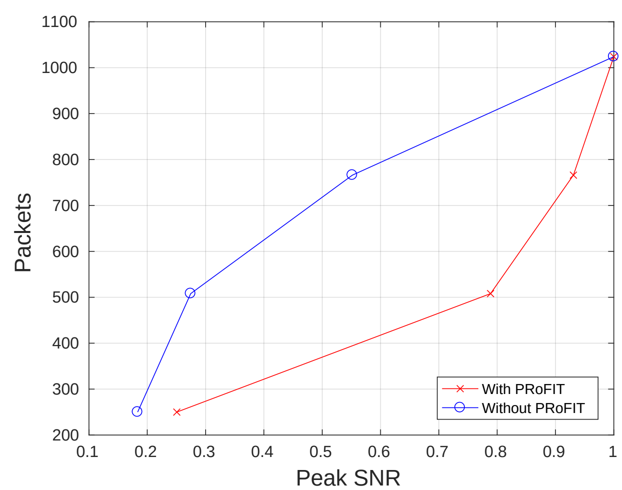


Fig. 20. Energy Consumption of image 256x256 Outdoor

Fig. 21 shows the relationship between number of packets received and accumulated PSNR of the *lena.tif (512 x 512)* image resolution. The image with PRoFIT achieved PSNR value of 0.8 when approximately 2,000 packets were received (48% of the total packets), wheres the image without PRoFIT reached PSNR value of 0.8 when 3,800 packets were received (93% of total packets). Therefore, we can conclude that the energy consumed in image transmission with PRoFIT is 45% less than the energy consumed in transmission without PRoFIT.

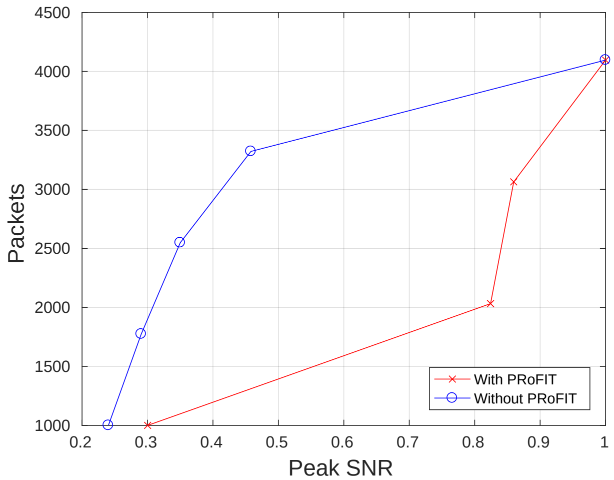


Fig. 21. Energy Consumption of image 512x512 Outdoor

VI. Conclusion and Future Research

This paper has provided the details of experiments done using our proposed framework called PRoFIT. PRoFIT is designed to deliver critical image features at high priority to the sink node for early processing. We built a real wirless node testbed to evaluate PRoFIT. Peak signal-to-noise ratio (PSNR) analysis show that PRoFIT improves VSN application response time as compared to priority-less routing. This paper also contains the design of our VSN testbed. Multiple indoor and outdoor experiments were performed to validate the framework. This framework also improves the energy efficiency of the network.

In future, we plan to enhance the encoding mechanism by introducing multilevel diversity coding (MLDC) and advance encoding for video transmission. The communication will be improved by enhancing the MAC layer of Zigbee. PRoFIT will be modified for special cases in video transmission. The testbed will be improved using GPUs, energy efficient components and software management tools. Many applications can be develop using PRoFIT and the real testbed.

Acknowledgments

This work is funded by the Grant Number AT-34-335, King Abdul Aziz City of Science and Technology (KACST) of the Kingdom of Saudi Arabia. The authors thank KACST for their support.

References

[1] L. Makkaoui, V. Lecuire, and J. M. Moureaux, *Fast zonal dct-based image compression for wireless camera sensor networks*, in Proc of the 2010 2nd International Conference on Image Processing Theory, Tools and Applications, July 2010, pp. 126-129.

[2] M. Nasri, A. Helali, H. Sghaier, and H. Maaref, *Energy-efficient wavelet image compression in wireless sensor network*, in Proc. of the 2010 International Conference on Wireless and Ubiquitous Systems, Oct 2010, pp. 1-7.

[3] V. Delport, J. Kuhnert, and S. Roessler, *Implementation of an ieee 802.15.4 compliant multi-hop wireless sensor network for energy efficient sensor data and progressive image transmission*, in Proc. of the 2014 6th European Embedded Design in Education and Research Conference (EDERC), Sept 2014, pp. 182-186.

[4] H. Kikuchi, K. Funahashi, and S. Muramatsu, *Simple bit-plane coding for lossless image compression and extended functionalities*, in Proc. of the 2009 Picture Coding Symposium, May 2009, pp. 1-4.

[5] E. Felemban, A. A. Sheikh, and M. A. Manzoor, *Improving response time in time critical visual sensor network applications*, Ad Hoc Networks, vol. 23, pp. 65-79, 2014. [Online]. Available:http://www.sciencedirect.com/science/article/pii/S1570870514001140.

[6] Adil A. Sheikh, Atif Naseer, Emad Felemban, *PRoFIT - Priority-based Routing Framework for Image Transmission in Visual Sensor Networks*, In proceedings of The 4th International Conference on Communications and Network Engineering, 2017

[7] F. Zhou, Z. Lu, J. Liu, Y. Xia and X. Wei, *Low-Cost Prioritization of Microorganism Image Data in Wireless Visual Sensor Network*, 2016 International Computer Symposium (ICS), Chiayi, 2016, pp. 685-689.

[8] G. Han, J. Jiang, M. Guizani and J. J. P. C. Rodrigues, *Green Routing Protocols for Wireless Multimedia Sensor Networks*, in IEEE Wireless Communications, vol. 23, no. 6, pp. 140-146, December 2016.

[9] L. Bondi, L. Baroffio, M. Cesana, A. Redondi and M. Tagliasacchi, *EZ-VSN: An Open-Source and Flexible Framework for Visual Sensor Networks*, in IEEE Internet of Things Journal, vol. 3, no. 5, pp. 767-778, Oct. 2016.

[10] E. Eriksson, G. Dan and V. Fodor, *Coordinating Distributed Algorithms for Feature Extraction Offloading in Multi-Camera Visual Sensor Networks*, in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 11, pp. 3288-3299, Nov. **2018**.

[11] V. D. Bhosale and R. A. Satao, *Lifetime maximization in mobile visual sensor network by priority assignment*, 2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), Kumaracoil, 2015, pp. 695-698.

[12] P. Spachos, A. K. Marnerides and D. Hatzinakos, *Content Relevance Opportunistic Routing for Wireless Multimedia Sensor Networks*, 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, 2014, pp. 263-268.

[13] P. Spachos, D. Toumpakaris and D. Hatzinakos, *QoS and energy-aware dynamic routing in Wireless Multimedia Sensor Networks*, 2015 IEEE International Conference on Communications (ICC), London, 2015, pp. 6935-6940.

[14] J. Horneber and A. Hergenroder, *Survey on testbeds and experimentation environments for wireless sensor network*, IEEE Communication Surveys and Tutorials, vol. 16, pp. 1820-1838, 2014.

[15] T. M. I. F. Akyildiz and K. R. Chowdhury, *Wireless multimedia sensor networks applications and testbeds*, proceedings of the IEEE, vol. 96, pp. 1588-1605, 2008.

[16] J. R. M.-D. D. Adrian Jimenez-Gonzalez, *An integrated testbed for cooperative perception with heterogeneous mobile and static sensors*, Sensors, vol. 11, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22247679

[17] R. W. Abdelhamid Mammeri, Azzedine Boukerche and D. Zhou, *Progressive compression and transmission of images experimental evaluation over visual sensor networks*, in Globecom 2013 - Ad Hoc and Sensor Networking Symposium, 2013, pp. 420-425..

[18] W. J.-W. Z. B. Cheng, L. Cui and P. H. Gerhard, *Multiple region of interest coverage in camera sensor networks for teleintensive care units*, IEEE Transaction on Industrial informatics, vol. 12, pp. 2331-2341, 2016.

[19] V. K. V. P. Munishwar and N. B. Abu-Ghazaleh, *Coverage in visual sensor networks with pan-tilt-zoom cameras: The maxfov problem*, in EEE INFOCOM 2014 - IEEE Conference on Computer Communications, Toronto, 2014, pp. 1492-1500.

[20] roseek.com, All-in-one smart cameras for traffic - 2018: Roseek Cheetah1 http://www.roseek.com/en/product/cheetah1.html (accessed Apr 29, 2018).

A Deep Learning Approach for Handwritten Arabic Names Recognition

Mohamed Elhafiz Mustafa[1], Murtada Khalafallah Elbashir[2]
College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia[1,2]
College of Computer Science and Information Technology, Sudan University of Science and Technology, Khartoum, Sudan[1]
Faculty of Mathematical and Computer Sciences, University of Gezira, Wad Madani, Sudan[2]

***Abstract*—Optical Character recognition (OCR) has enabled many applications as it has attained high accuracy for all printing documents and also for handwriting of many languages. However, the state-of-the-art accuracy of Arabic handwritten word recognition is far behind. Arabic script is cursive (both printed and handwritten). Therefore, traditionally Arabic recognition systems segment a word to characters first before recognizing its characters. Arabic word segmentation is very difficult because Arabic letters contain many dots. Moreover, Arabic letters are context sensitive and some letters overlapped vertically. A holistic recognizer that recognizes common words directly (without segmentation) seems the plausible model for recognizing Arabic common words. This paper presents the result of training a Conventional Neural Network (CNN), holistically, to recognize Arabic names. Experiments result shows that the proposed CNN is distinct and significantly superior to other recognizers that were used with the same dataset.**

Keywords—Deep learning; Arabic names recognition; holistic paradigm

I. Introduction

Arabic is the official language of more than twenty countries and the mother tongue of more than 300 million people [6]. Arabic is one of the six United Nations official languages. The Arabic script is also used as a medium of writing for other languages like Persian. Moreover, Arabic script is the former script (Ottoman script) of the Turkish language [23]. Unfortunately, research in Arabic language recognition is far behind when compared with languages with the same size and importance. For instance, number of big popular Arabic datasets that are widely used is very few [14]. Moreover, most of these datasets are small. For instance, the IFN/ENIT dataset which is widely used by researchers and also used in ICDAR Arabic recognition competition contains only 26559 word images (names of Tunisian cities) [17], [22].

Arabic script is cursive (both handwritten and machine-printed text). Fig. 1, presents the Arabic alphabet and their different shapes. This makes handwriting recognition very challenging, as we need word segmentation before applying character recognition system. The segmentation methods reported in the literature are far from being robust and very accurate [1], [5]. Segmenting Arabic words to characters is very difficult [12]. There are two reasons for this difficulty. First Arabic letter shape is context sensitive. Some Arabic letters have four shapes according to their position in the word (see Fig. 1). Secondly, in Arabic writing dots are very important and they are not few as fifteen letters out of twenty eight have dots above or below them. Arabic writers do not place these dots carefully on their proper place and this leads to much confusion. although Arabic writing is horizontal from right to left some letters overlapped vertical. For illustration Fig. 2, contains an image of the Arabic word " حجج " (arguments in English). The image shows three different handwriting. For comparison the Figure also shows some printing for the same word. This is an example for a word that expert Arabic reader recognizes it as one unit without trying to figure out each letter and its dot(s). An analytical automatic recognition system should find the proper segmentation first and then it should find the proper coupling of the dots with its corresponding letter.

In the literature, there are good results for isolated Arabic character recognition (in some papers the recognition is more than 97%) [21], [4], [2], [3]. However, publication on word recognition are few with low recognition accuracy rate [15], [16]. This low rate for word recognition accuracy is mainly due to the error in segmentation [1]. Arabic readers generally tend to recognize common Arabic words and names holistically (i.e. without segmenting them). For new words or non-common ones, the Arabic readers identify the word letters and then recognize the word. This idea becomes appealing for automatic recognition and found good support in last couple of decades [10]. For Arabic words, holistic paradigm is faster and, in some cases, more accurate than analytic paradigm [19], [11]. The disadvantage of the holistic paradigm is its need for huge training data. Moreover, the holistic paradigm does not work alone as the recognition system needs to revert to the analytic paradigm when the word under consideration is not in common words lexicon. bypassing the problems that occur as a result of over segmentation and under segmentation is the main advantage of holistic approach-based classifiers for OCR.

Arabic names used today have much repetition. For instance, in SUST-ARG dataset [13], the twenty most frequent male names represent 52% of the total male names. Application forms normally contain the applicant name. Therefore, for the automatic processing of these forms, the holistic paradigm seems appropriate to quickly recognize common names and to resort to the analytical paradigm to recognize uncommon ones.

In recent years, deep learning has gained great popularity in the pattern recognition filed [9]. It becomes the focus of many researchers since it represents the easiest way to deal with huge data and it automates the feature extraction task. From our literature survey, we notice that although deep learning is the state of the art for OCR, the number of papers which use deep learning for Arabic character recognition is very few [21], [4], [2], [8]. Moreover, All these work address isolated

alone	initial	medial	final	name
ا	ا	ـا	ـا	alif
ب	بـ	ـبـ	ـب	bāʾ
ت	تـ	ـتـ	ـت	tāʾ
ث	ثـ	ـثـ	ـث	thāʾ
ج	جـ	ـجـ	ـج	jīm
ح	حـ	ـحـ	ـح	ḥāʾ
خ	خـ	ـخـ	ـخ	khāʾ
د	د	ـد	ـد	dāl
ذ	ذ	ـذ	ـذ	dhāl
ر	ر	ـر	ـر	rāʾ
ز	ز	ـز	ـز	zāʾ
س	سـ	ـسـ	ـس	sīn
ش	شـ	ـشـ	ـش	shīn
ص	صـ	ـصـ	ـص	ṣād
ض	ضـ	ـضـ	ـض	ḍād
ط	طـ	ـطـ	ـط	ṭāʾ
ظ	ظـ	ـظـ	ـظ	ẓāʾ
ع	عـ	ـعـ	ـع	ʿayn
غ	غـ	ـغـ	ـغ	ghayn
ف	فـ	ـفـ	ـف	fāʾ
ق	قـ	ـقـ	ـق	qāf
ك	كـ	ـكـ	ـك	kāf
ل	لـ	ـلـ	ـل	lām
م	مـ	ـمـ	ـم	mīm
ن	نـ	ـنـ	ـن	nūn
ه	هـ	ـهـ	ـه	hāʾ
و	و	ـو	ـو	wāw
ي	يـ	ـيـ	ـي	yāʾ

Fig. 1. Arabic alphabet and their different shapes

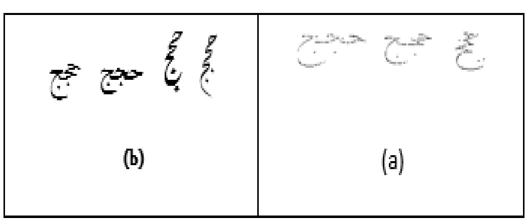


Fig. 2. This figure shows different handwriting and printing for a single Arabic word

Arabic letters and numerals recognition. Therefore, there is a lack of research that uses deep learning to recognize Arabic words holistically (i.e. using holistic paradigm).

The rest of the paper is organized as follows: The next section describes the materials and methods. The material and methods starts by presenting SUST-ARG names dataset

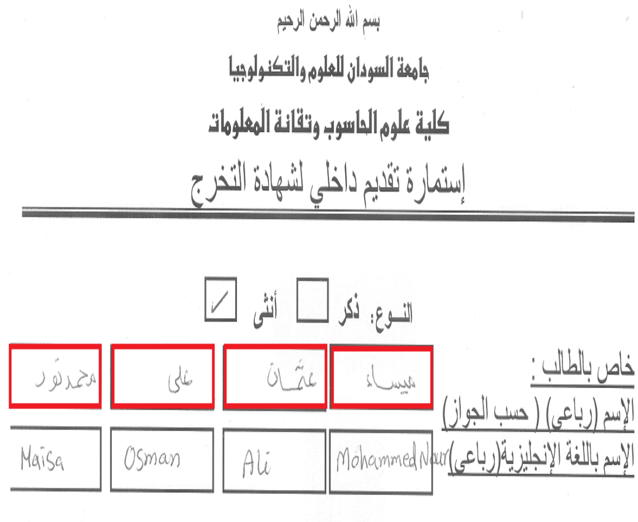


Fig. 3. Upper part of SUST certificate application form, extracted names are indicated by the red boxes.

followed by describing the preprocessing of the data, and finally describes the proposed model. Section 3, presents and discusses the results. the conclusions and future works are given in Section 4.

II. Material and Methods

A. Dataset: SUST-ARG names

The Dataset used in this paper, SUST-ARG names, is collected by the Arabic Recognition Research Group in the College of Computer Science and Information Technology, Sudan University of Science and Technology. This dataset has been collected in two stages. In the first stage 8028 names written by 2007 students have been extracted from the Certificate application form which used by the registrar office in the college (Fig. 3). From this data the most repeated names have been identified. In the second stage, a new form is designed to collect data for Forty common names (Fig. 4). The right column in Figure 4 contains male names and the left for female names. One thousand forms have been filled by university students. The forms were preprocessed and a dataset that contains forty thousand names is constructed. This dataset contains two parts, twenty thousand male names dataset and twenty thousand female names. Only male names are used in the experiments of this paper. Table I, contains male names samples.

B. Data Preprocessing

The data collection forms were digitized using a scanner and then the names are extracted from the scanned paper. Most of the names are not in the center of their boxes as depicted in Table I. The initial experiments show that raw data is not suitable as input to the proposed CNN. The extracted images are then preprocessed by eliminating the surrounding white spaces and down scaling the image to 28 x 56 pixels. After that all the images are converted to black Background and white foreground.

Fig. 4. Second stage data collection form sample.

TABLE I. MALE NAMES SAMPLES.

#	English name	Arabic name	Samples
1.	Ibrahim	إبراهيم	
2.	Ahmed	احمد	
3.	Idris	إدريس	
4.	Adam	آدم	
5.	Ismail	إسماعيل	
6.	Babiker	بابكر	
7.	Bashir	بشير	
8.	Gafar	جعفر	
9.	Hamid	حامد	
10.	Hassan	حسن	
11.	Hussain	حسين	
12.	Khalid	خالد	
13.	Saeed	سعيد	
14.	Soliman	سليمان	
15.	Salih	صالح	
16.	Siddiq	صديق	
17.	Taha	طه	
18.	Abass	عباس	
19.	Osman	عثمان	
20.	Ali	علي	

C. The Proposed CNN Architecture

Convolutional Neural Networks (CNNs) is one of the important classes of deep learning, which is mostly applied in computer vision to analyze and identify visual imagery [20]. They are found to be very efficient in computer vision problem and they have state of the art performance when provided with large amount of training data [18]. Normally, a CNN architecture consists of a sequence of layers that are stacked in specific organization where the output of each layer is used as an input for the layer that comes after it. The basic layers that can be found in any CNN architecture are: convolutional layer, pooling layer, and the fully connected or dense layer. The most important layer in the CNN architecture is the convolutional layer and normally it is the first layer (the input layer). It uses what is known as convolutional filtering to detect the presence of features that can be used to distinguish specific image that is provided as an input by creating features map. These feature map records the exact position of features in the image, where small movements in this position will result in a totally different feature map. Down sampling is one of the approaches to address the problem that can be created by the features map, and it is implemented in the CNN architecture by the means of a max pooling layer, which scales the number of features generated by the convolutional layer and keep the most important features. There are two important functions that can be used in the pooling layer: Max pooling and Average pooling, where the former takes the large value in each patch of features i.e. the large value in a set of pixels and the later calculates the average of each patch. Because the distribution of each layer input changes with the change of its previous layer parameters during the training process, a batch normalization is needed to normalize layer inputs and hence speed up the learning process [7]. The layer that is used to classify the input images is known as the fully connected layer and normally it is the last layer in the CNN architecture. It produces outputs equivalent to the number of the classes in the classification problem. These outputs are probabilities for the class's labels. Activation functions are essential part of the CNN architecture to accomplish the complex functional mapping between the input and the outputs, which can be used as inputs for the next layer. The proposed CNN architecture model for Arabic names recognition uses four convolutional layers and each layer uses Rectified linear unit (RelU) activation function, which is very popular activation function. According to the literature, it has very good experimental results. Simply it is defined as shown in equation 1.

$$f(x) = max(0.x) \tag{1}$$

The function will return zero if the value is negative and the value itself if it is positive, which means that it grows linearly for positive values. Because Max pooling layer can select the brighter pixels from the inputs and it suits the case where the background is dark i.e. the same as the preprocessed images in our dataset, we used it after the second and the fourth convolutional layers. Our proposed model uses two dense layers and it uses softmax activation function after the second Dense layer. The softmax function forces the output of each unit to be between 0 and 1. It divides each output such that the total sum of the outputs is equal to 1 and mathematically can be expressed as shown in equation 2.

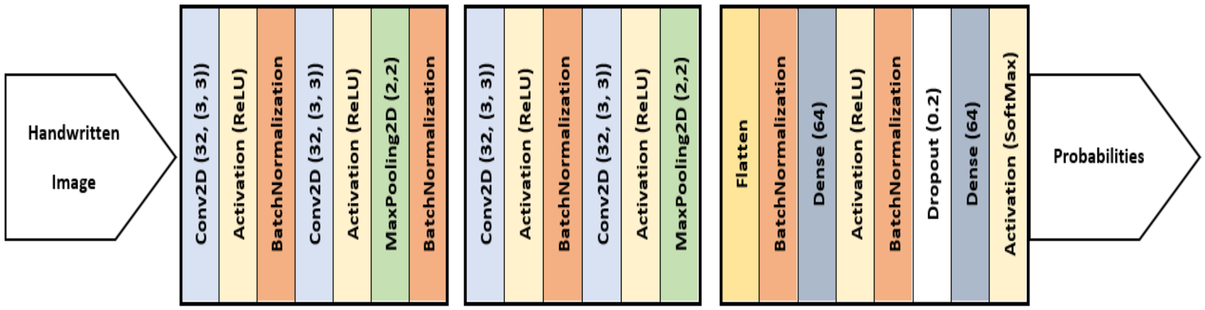


Fig. 5. The architecture of the proposed model

$$\sigma(Z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K}(e^{z_k})} \quad (2)$$

where z is a vector of the inputs. j is an index for the output units and it goes from 1 to K.The complete architecture of our proposed model is depicted in Fig. 5.

III. Results and Discussion

We used keras library under TensorFlow environment to construct the proposed CNN model. Ten fold cross validation is used to train and test the proposed model, where the entire dataset is divided into 10 folds. One of these 10 folds is used as a testing set and the remaining folds are used for training the proposed model. This process is repeated 10 times by taking a unique fold in each time. The final results is calculated as the average of the results that are obtained from the repetition. The accuracy of prediction is used as performance measure. The accuracy obtained for the testing names is 99.14%. Table II shows some of the missed classified names.

The literature survey reveals that there are two papers which describe recognition models using the same data set [19], [11]. Both of the papers have used only the male names part of the dataset (same as the experiments of this paper). The classifier of the first paper is HMM and the best classification accuracy is 63%. The classifier of the second paper is probabilistic neural networks, where the classification accuracy is 89% (see Table III).

Despite the fact that the classification accuracy of the proposed model is very high, however, one may consider that the small number of classes (only 20 names) is a major drawback. As we have mentioned holistic classifier does not work alone. The classifier will recognize common names only and the system should use the analytical approach to recognize the new names or uncommon ones. Therefore, there is critical trade-off here, increasing the number of names, speeds up the processing but the cost of training is higher. Moreover, increase the number of classes may decrease the accuracy. This is a question for further research.

IV. Conclusion

This paper tests the usage of Holistic paradigm to recognize Arabic names without segmentation. As deep learning is currently the state of the art for such system, a CNN model is designed, trained and tested using SUST-ARG-names dataset. The goal of the paper is to study the idea of building a system which recognizes common Arabic names very quickly. Although the accuracy is very high, the number of names selected for the experiments of this paper is relatively small (20

TABLE II. Sample of missed classified names.

#	English name	Arabic name	Sample	Classified as
1	Ibrahim	إبراهيم		Ismail
2	Ahmed	احمد		Hassan
3	Idris	إدريس		Ibrahim
4	Adam	آدم		Idris
5	Ismail	إسماعيل		Osman
6	Babiker	بابكر		Ahmed
7	Bashir	بشير		Hussain
8	Gafar	جعفر		Saeed
9	Hamid	حامد		Khalid
10	Hassan	حسن		Hussain
11	Hussain	حسين		Hassan
12	Khalid	خالد		Hamid
13	Saeed	سعيد		Bashir
14	Soliman	سليمان		Hussain
15	Salih	صالح		Khalid
16	Siddiq	صديق		Saeed
17	Taha	طه		Khalid
18	Abass	عباس		Osman
19	Osman	عثمان		Soliman
20	Ali	علي		Khalid

TABLE III. This table shows the accuracy of the proposed classifier together with the accuracy of two classifiers trained on the same dataset.

Classifier	Accuracy
The proposed model	99.14%
Hidden Markov Model [19]	63%
Probabilistic neural network [11]	89%

names). Comparing the proposed model with previous similar work on the same dataset is the main reason for selecting this small number. According to the statistics on the given dataset, these 20 names represent more than 50% of SUST students' names. It seems clear that increasing the number of names to represent, for instance, more than 90% of the used name is the future work that must be done to support and prove the practicality of this model.

REFERENCES

[1] Yasser M Alginahi. A survey on arabic character segmentation. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(2):105–126, 2013.

[2] Chaouki Boufenar and Mohamed Batouche. Investigation on deep learning for off-line handwritten arabic character recognition using theano research platform. In *2017 Intelligent Systems and Computer Vision (ISCV)*, pages 1–6. IEEE, 2017.

[3] Murtada Khalafallah Elbashir and Mohamed Elhafiz Mustafa. Convolutional neural network model for arabic handwritten characters recognition. *International Journal of Advanced Research in Computer and Communication Engineering*, 7(11), 2018.

[4] Mohamed Elleuch, Najiba Tagougui, and Monji Kherallah. Arabic handwritten characters recognition using deep belief neural networks. In *2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD15)*, pages 1–5. IEEE, 2015.

[5] Hashem Ghaleb, P Nagabhushan, and Umapada Pal. Segmentation of overlapped handwritten arabic sub-words. *International Journal of Computer Applications*, 975:8887, 2015.

[6] Barbara F Grimes. Ethnologue: Languages of the world. 14. *Aufl. Dallas*, 2000.

[7] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[8] Ahmed Lawgali. A survey on arabic character recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(2):401–426, 2015.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[10] Sriganesh Madhvanath and Venu Govindaraju. The role of holistic paradigms in handwritten word recognition. *IEEE transactions on pattern analysis and machine intelligence*, 23(2):149–164, 2001.

[11] Wafa A. Mohamed and M. E. M. Musa. Recognition of arabic handwritten names using probabilistic neural networks. *Computer Studies Journal*, 1(1), 2009.

[12] Bodoor Bashir Mohamed E. M. Must and Mohamed N. I. Ismail. Designing an arabic handwritten segmentation system. *Intenational Journal of Computing*.

[13] Mohamed EM Musa. Towards building standard datasets for arabic recognition. *International Journal of Engineering and Advanced Research Technology (IJEART)*, 2(2), 2016.

[14] Mohammad Tanvir Parvez and Sabri A Mahmoud. Offline arabic handwritten text recognition: a survey. *ACM Computing Surveys (CSUR)*, 45(2):23, 2013.

[15] Mohammad Tanvir Parvez and Sabri A Mahmoud. Offline arabic handwritten text recognition: a survey. *ACM Computing Surveys (CSUR)*, 45(2):23, 2013.

[16] MS Patel and Rohith Kumar. Offline handwritten word recognition: A survey.

[17] Mario Pechwitz, S Snoussi Maddouri, Volker Märgner, Noureddine Ellouze, Hamid Amiri, et al. Ifn/enit-database of handwritten arabic words. In *Proc. of CIFED*, volume 2, pages 127–136. Citeseer, 2002.

[18] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[19] Talaat M. Wahbi and Mohamed E. M. Musa. Holistic approach for arabic word recognition. *international Journal of Computer Applications Technology and Research*, 5(3):141–146, 2016.

[20] Wenzhe Wang, Bin Wu, Sixiong Yang, and Zhixiang Wang. Road damage detection and classification with faster r-cnn. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5220–5223. IEEE, 2018.

[21] Khaled S Younis. Arabic handwritten character recognition based on deep convolutional neural networks. *Jordanian Journal of Computers and Information Technology*, 3(3), 2017.

[22] Oussama Zayene, Jean Hennebert, Rolf Ingold, and Najoua Essoukri BenAmara. Icdar2017 competition on arabic text detection and recognition in multi-resolution video frames. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1460–1465. IEEE, 2017.

[23] Erik J Zürcher. *Turkey: A modern history*. Bloomsbury Publishing, 2017.

Optimal Topology Generation for Linear Wireless Sensor Networks based on Genetic Algorithm

Adil A. Sheikh[1]
Science and Technology Unit
Umm Al-Qura University
Makkah, Saudi Arabia

Emad Felemban[2]
College of Computing and Information Systems
Umm Al-Qura University
Makkah Saudi Arabia

***Abstract*—A linear network is a type of wireless sensor network in which sparse nodes are deployed along a virtual line; for example, on streetlights or columns of a bridge, tunnel, and pipelines. The typical deployment of Linear Wireless Sensor Network (LWSN) creates an energy hole around the sink node since nodes near the sink nodes deplete their energy faster than others. Optimal network topology is one of the key factors that can help improve LWSN performance and lifetime. Finding optimal topology becomes tough in large network where total possible combinations is very high. We propose an Optimal Topology Generation (OpToGen) framework based on genetic algorithm for LWSN. Network deployment tools can use OpToGen to configure and deploy LWSNs. Through a discrete event simulator, we demonstrate that the use of genetic algorithm accomplishes fast convergence to optimal topologies as well as less computational overhead as compared to brute force search for optimal topology. We have evaluated OpToGen on the number of generations it took to achieve the best topology for various sized LWSNs. The trade-off between energy consumption and different network sizes is also reported.**

***Keywords*—Ad hoc networks; network topology; genetic algorithms; computer simulation; computer networks management; network lifetime estimation; optimization**

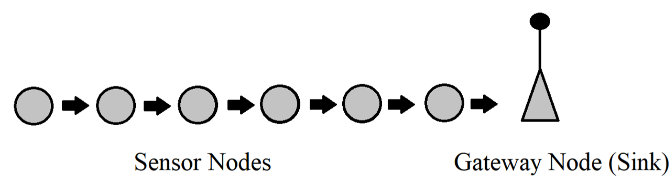


Fig. 1. Linear Wireless Sensor Network

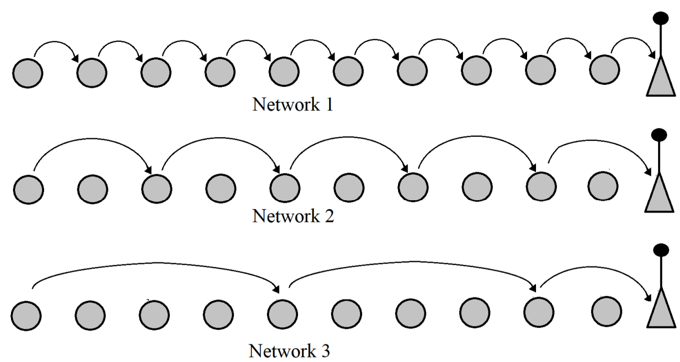


Fig. 2. Typical Connectivity Scenarios of LWSNs

I. Introduction

Linear Wireless Sensor Network (LWSN) is a special kind of wireless sensor networks where nodes are placed in a linear formation (Fig. 1). LWSNs are used in applications involves monitoring or collecting data from infrastructure that spans over long distances such as pipelines, highways, bridges, and borders. Usually, WSN is deployed in 2D or 3D form where nodes are connected with multiple nodes in all directions. LWSN, by definition, on the other hand, have minimal connectivity. Fig. 2, shows typical connectivity scenarios for a linear network.

Typically, LWSN does not need sophisticated routing protocols since nodes generate packets and forward them to the next node closer to the sink node. As a result, nodes closer to the sink node have a higher load and thus deplete their batteries faster than others creating an energy hole around the sink node. There have been enormous efforts [1], [2], [3], [4] to characterize and study this kind of networks, including throughput improvement, lifetime enhancement, end-to-end delay minimization, and others. Topology configuration in a linear network is minimal; however, it is imperative since performance is much dependent on it [4].

Topology optimization in WSN has been an active research area recently [5] where selecting topology that satisfies specific criteria is the objective. Approaches such as Ant Colony Optimization, Swarm Optimization, Neural Networks, Genetic Algorithms, and Artificial Intelligence have been presented to solve the topology optimization problem [6]. Furthermore, the optimization of simple deployment like LWSNs would become complex when many parameters affect the outcome. The use of LWSN in smart cities' intelligent transportation systems has pushed researchers to refocus on finding optimal topologies. Our proposed framework leverages the advancements in genetic algorithms to address complex problems, such as increasing network lifetime by optimizing network topology.

In our previous work [7], we introduced OpToGen concept. In this paper, we propose OpToGen framework that uses GA in conjunction with a network simulator to optimally deploy a heterogeneous LWSN. The optimization module that uses GA to maximize network lifetime is the main contribution of this paper. We constructed a fitness function based on network parameters related to network lifetime. Using GA has resulted in significantly reducing the computational overhead in finding the best or near-best network topology. We validate the practicality of OpToGen by implementing the framework on Matlab and NS-2. In future works, the researchers would like to address limitations of OpToGen, such as scalability.

The rest of the paper is divided into the following sections. The next section, Section II, discusses related research. Section III discusses OpToGen framework and its design. The implementation of OpToGen framework as a tool and its evaluation are presented in Section IV. The paper is summarized in Section V. Finally, possible future directions are presented in Section VI.

II. Related Work

In the past decade, there has been a significant advancement in research on various issues related to LWSNs. Topological optimization of networks is of major importance due to its requirement in many domains, including telecommunications, electricity distribution, underwater cables, and oil/gas pipelines. Topology optimization can solve challenging problems like network lifetime maximization, node coverage, energy consumption, security, and reliability.

The author in [8] provides a comprehensive overview of linear wireless sensor networks (LWSNs) including its concept, applications and motivation to design specific network protocols that explores linearity of a network for energy savings, increasing lifetime, fault tolerance and reliability. A detailed hierarchical and topological classification of LWSN is also provided. A case study based on associated protocols is presented to increase the robustness and efficiency of routing for network optimization.

An efficient and reliable LWSN design is presented by [9]. In their design, the requirement of energy consumption and long network lifetime are met using an optimal node deployment scheme for data flow. Network efficiency is the primary optimization objective. The node deployment scheme optimizes the number of clusters and the number of relay and sensor nodes to be considered to form a linear network. Results of theoretical analysis and simulations for their work show that the method of nodes placement and topology control can solve energy consumption problems and improve network efficiency.

In [10], the author provides an analytical framework for node placement in a linear array fashion for uniform energy dissipation of all sensor nodes. This approach makes sure that all sensor nodes in the network die out simultaneously. The results show that the network lifetime is doubled as compared to other mentioned approaches. The author has also mentioned issues related to random node placement, as this theoretical requirement is not fulfilled in real scenarios. The random node placement with fixed bin length outperforms random node placement with variable bin length.

An efficient node placement scheme is presented in [11] to maximize the lifetime of LWSN through uniform energy conservation. The performance of uniform and random and linearly decreasing distance (LDD) node placement schemes are evaluated. The impact of gateway location on network performance is also analyzed. In a random node placement scheme, the nodes are placed randomly. For a connected network, it is assumed that each node is in communication range of the previous one. In uniform node placement scheme, the nodes are placed at equal distance with GW at the edge. In this case, nodes near GW lose energy soon due to additional data from remote sensors. In LDD, the distance between nodes is decreased toward GW to balance the load and increase the lifetime of the network. All these approaches result in increased network lifetime when GW is placed in the middle due to the reduced overhead of data forwarding in nearby nodes.

In literature, topological optimization problems can also be solved by heuristic methods viz. simulated annealing [12], Artificial Neural Network [13] and GA [14], [15] are preferred because of their strength and ability to find near-optimal solutions. Researchers have used GA to optimize different parameters of various network topologies. A survey is provided in [16] based on the application of GAs for sensor network optimization. The authors provide a comprehensive review modeling sensor communication in clustering and routing problems using GA. They also evaluate various GA-based optimization strategies.

In [17], the authors present an energy efficient network layout based on GA. The objectives taken into consideration are coverage and lifetime for network reliability. A novel method is introduced to decrease communication energy by clustering nodes and positioning them in the closest possible distance in each cluster. In their work, nodes are positioned in the network using GA and nodes have been divided into specific clusters using a K-means clustering algorithm. Another effort is reported by [18] where an algorithm is introduced to maximize the reliability of the network. In their algorithm, a different and efficient chromosome, as well as gene encoding and cross over, is used for better convergence towards the optimal solution. The population size and the computational time of the networks ensure that the proposed method converges to its optimal solution in very few CPU seconds. An efficient topology control using GA based cluster head selection is presented by [19].

In their clustering technique, the 3-tier sensor network architecture is developed comprising of super head nodes, cluster head (CH) nodes, and cluster slave nodes. This architecture utilizes node's residual energy, bandwidth, and memory capacity to reduce the energy consumption of the network. Research work of [20] presents an algorithm to optimize the design of communication network topologies that are based on GA. A set of links for a given set of nodes is chosen to maximize reliability. Another complication that their algorithm solves is the calculation of computationally expensive reliability fitness function for promising candidate networks.

The study of related works mentioned above suggests that most research works have been carried out to optimize topologies for improving the reliability of WSNs. Some researchers have directed their research towards finding out optimal clustering of nodes to reduce communication overhead, gene and chromosome encoding to reduce computation time, node deployment scheme for energy efficiency and online interactive algorithm to maximize network lifetime. Others have provided ways to validate optimal topologies through various means. The work done in this paper tries to determine the best or near-best topology in lesser time and computation overhead using a novel GA under some constraints set on LWSNs.

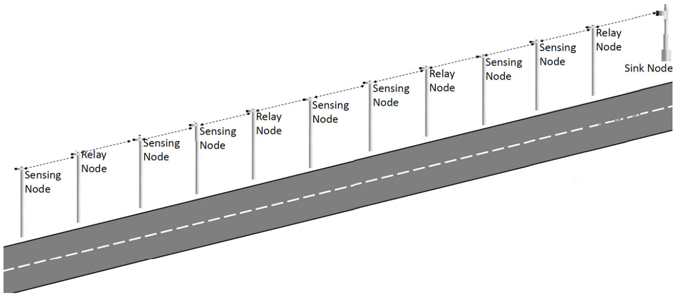


Fig. 3. Sample LWSN Scenario

III. OpToGen Framework

In this research, we focus on two-tiered LWSN however, OpToGen generates optimal topologies for multi-tiered LWSNs. Fig. 3 depicts the scenario under discussion where sensor nodes or cameras have been attached to light poles along a highway. A sink node that collects and process data from sensor and camera node is placed in the network where it is in the direct transmission range of only a few nodes. Therefore, each sensor node relays its collected data through relay nodes to the sink node.

In this way, sensor nodes with limited memory, low processing power, and low battery power are part of tier-2 from the sink, whereas, the higher powered relay node is part of tier-1. Relay nodes are not limited to forwarding data from other nodes. They can also take part in sensing.

The scenario described above is given as input to OpToGen. OpToGen's role as a black box is shown in Fig. 4. The number of nodes and other network parameters is given as input to OpToGen framework. Using GA and network simulator, OpToGen takes significantly less time to generate optimal or near-best topology. This is explained in Section III of the paper.

Finding optimal topology for a dense LWSN has been actively researched for some time now, but using network simulations to evaluate the fitness function is a novel idea. In [21], the researchers show a classic decision-making process based on GA. Fig. 5a explains their work. The first step is to create an initial population of solutions. The second step is to assess the fitness of that population. No further action is required if the desired fitness is achieved. Otherwise, the breeding process is carried out using genetic operators to create a new generation. Next, the new generation is evaluated. This evolution cycle continues until fitness no longer improves. We have modified the classic decision-making process of GA. As depicted in Fig. 5b, a network simulator is introduced for evaluating the current population's fitness.

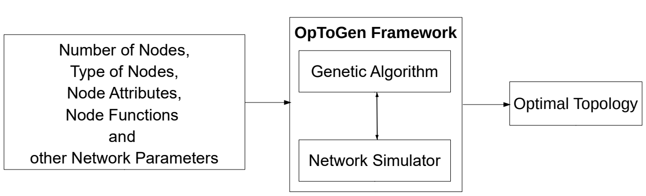
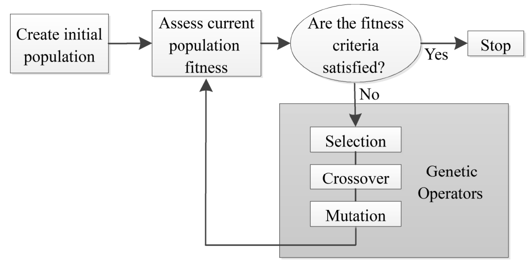
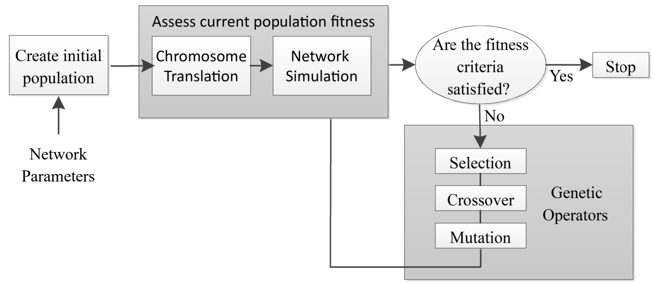


Fig. 4. OpToGen as a Black Box



(a) Typical GA based Decision Making Process [20]



(b) OpToGen Framework Decision Making Process

Fig. 5. Typical GA vs. OpToGen

A. OpToGen's Genetic Algorithm

Evolutionary Computation (EC) is the field of study that focuses on the simulation of natural evolution as a genetic solver for intractable problems and optimizations. OpToGen is based on EC. To achieve higher network lifetime, GA determines the optimal topology by determining optimal connections between LWSN nodes.

The working of OpToGen framework is explained using Algorithm 1. The inputs to OpToGen include *N*, the total number of sensor nodes to be deployed. Note that for a structure like a highway, the architect would know the total light poles on the highway. The remaining inputs are for GA. These include *G*, the maximum number of generations that the GA should run. P_C is the probability of crossover and P_M is the probability of mutation. The ratio of parents to offsprings in the new generation is *R*. Finally, *I* is the number of individuals in each population.

Like any other EC, OpToGen starts with generating its primary population. A random population of candidate topologies or candidate chromosomes is generated by *Generate_Random_Population()* at the start of the algorithm. The next sub-section explains how a chromosome represents a candidate topology.

B. Chromosome Representation

The first step to use GA for problem-solving is to design an appropriate encoding of the problem's candidate solution. This design problem in itself is challenging. A better candidate solution representation as a chromosome is essential for the efficiency of GA. This step includes both genotypes and phenotypes for the GA.

Algorithm 1: Pseudocode for Generating Optimal Topology

Function OpToGen (G, N, I, R, P_C, P_M);
Input : G is the maximum number of generations that the algorithm should run.
I is the number of individuals in each generation.
N is the number of nodes in the network.
R is the ratio of offsprings chromosomes to parent chromosomes in the new generation.
P_C and P_M are the probability of crossover and mutation, respectively.
Output: $Best\_Topology$ is the optimal network topology and $Best\_Fitness$ its fitness.
$G_1 \leftarrow Generate\_Random\_Population(N)$
$i \leftarrow 1$;
$Fitnesses \leftarrow \emptyset$
repeat
  $Topology_i \leftarrow Translate\_Chromosome\,(G_1(i))$
  $Fitness_i \leftarrow Network\_Simulator(Topology_i)$
  $Fitnesses \leftarrow (Fitnesses \cup Fitness_i)$
  $i \leftarrow i + 1$;
until $i \leq I$;
$g \leftarrow 1$;
repeat
  $i \leftarrow 1$;
  $Offsprings \leftarrow \emptyset$
  repeat
    $parent_1 \leftarrow Select\_Random\_Parent(G_g)$
    $parent_2 \leftarrow Select\_Random\_Parent(G_g)$
    $p_c \leftarrow Random\_Number()$
    if $p_c \leq P_C$ **then**
      $offspring_i \leftarrow Crossover(parent_1, parent_2)$
    else
      $offspring_i \leftarrow Select\_Random\_Parent(parent_1, parent_2)$
    $p_m \leftarrow Random\_Number()$
    if $p_m \leq P_M$ **then**
      $offspring_i \leftarrow Mutate(offspring_i)$
    $Topology_i \leftarrow Translate\_Chromosome\,(offsprring_i)$
    $Fitness_i \leftarrow Network\_Simulator(Topology_i)$
    $Fitnesses \leftarrow (Fitnesses \cup Fitness_i)$
    $Offsprings \leftarrow (Offsprings \cup offspring_i)$
    $i \leftarrow i + 1$;
  until $i \leq I$;
  $P \leftarrow Sort\_Descending(G_g, Fitness(G_g))$
  $O \leftarrow Sort\_Descending(Offsprings, Fitness(Offsprings))$
  $G_{g+1} \leftarrow \left(O_1, O_2, ..., O_{I.R}, P_1, P_2, ..., P_{I.(1-R)}\right)$
  $Best\_Topology \leftarrow O_1$;
  $Best\_Fitness \leftarrow Fitness(O_1)$
  $g \leftarrow g + 1$
until $g \leq G$;

A string representation of the candidate topology is known as genotype. The decoded representation of the string is known as the phenotype. Note that a candidate topology's genotype that is represented as a chromosome consists of, for example, a bit string — all the variations of the bits in this bit-string of that genotype cover the entire solution set. The search for optimal topology LWSN becomes a complex and intractable problem because of the high number of nodes and their associated parameters. Finding the optimal candidate would need a large number of computations. Therefore, GA based heuristics solution can be valuable in mitigating this problem.

As discussed above, for OpToGen, a chromosome represents an LWSN topology. The chromosome has been designed to represent sensor nodes (tier-2) as indices of the bit-codes. The bit values at each index represent either the sink node or relay node (tier-1). The architect is assumed to know the physical length of the LWSN beforehand, for instance, if the architect is planning to cover a highway or certain length with an LWSN of the camera and other sensor nodes. The architect would also know the field of view of the camera and detection ranges of the sensor nodes. Therefore, the architect would know the number of cameras and nodes required to cover the highway. The number of nodes, *N*, is one of the inputs of OpToGen. If we assume that the architect places one relay node between every two pairs of sensor nodes for maximum reliability, the number of relay node can be calculated to be the maximum of half of *N*. These values, along with the user-selected probability of crossover, P_C, and the probability of mutation, P_M, are given as input to OpToGen. Hence, we can list the assumptions as:

1) Sensor nodes can be less than or equal to twice the number of relay nodes.
2) Data can be sent directly to the sink node by a sensor node if it is in the transmission range.
3) Sensor nodes forward their data through a relay node if the sensor node is not in range of the sink node.
4) All relay nodes are powerful enough to be directly connected to the sink.

Let us take the example of an LWSN with *N = 6*. Assuming that the number of relay nodes is a maximum half of *N*, so we get three relay nodes and one sink. Therefore, our sample chromosome represents a network with one sink, three relay nodes, and six sensor nodes. Bit-string that represents this network has a length that can be calculated using Eq. 1.

$$N \times \left(\lceil log_2 \left(1 + \frac{N}{2} \right) \rceil \right) \tag{1}$$

From our example of a network of *N = 6*, Eq. 1 calculates bit-string length as 12-bits, where the indices of the bit pairs will represent the sensor node number and the value of the bits will represent the connection. This means that *bit-1* and *bit-2* represent sensor node 1's connectivity; *bit-3* and *bit-4* represent sensor node 2's connectivity and so on. The connectivity is determined by Eq. 2.

$$connection \equiv base_{10}(bitcode) \mod \left(1 + \lceil \frac{N}{2} \rceil \right) \tag{2}$$

Let us elaborate on the use of Eq. 2 with the help of Table I. Assuming the bit string has the value of *01111101000*, we break it down to sets of two bits each.

The first row of Table I contains indices of individual bits of the representing chromosome. The values of the bits are in the second row. The third row contains the ID of a sensor node whose connectivity is being depicted by the bits. Finally, the decoded connection, according to Eq. 2, is shown in the fourth row.

OpToGen designates the connection value of 0 to sink, the connection value of 1 to *Relay 1*, the connection value of 2 to *Relay 2* and so on. Therefore, using Eq. 2, OpToGen decoded bit code *00* as sink i.e. $(base_{10}(00)_2) \mod 4 \equiv 0$. Bit code 01 is decoded as Relay 1, i.e. $(base_{10}(01)_2) \mod 4 \equiv 1$. Similarly, bit code 10 is decoded as Relay 2 and bit code 11 as Relay 3. Therefore, the network represented by the chromosome sample in Table 1 has Sensor Node 1 and Sensor Node 2 connected to Relay 1. Sensor Node 3 connected to Relay 3. Sensor Node 4 and Sensor Node 5 are connected to Relay 2. Lastly, Sensor Node 6 is directly connected to the sink. This network is logically depicted in Fig. 6. Note that the actual physical network may be different. Also, note that all relay nodes are assumed to be connected to the sink; therefore, their representation in the chromosome is not required.

Let us take the example of another network, where N = 9. Using Eq. 1, the bit-string length is calculated to be *27*. Table II depicts a sample chromosome of length *27*. Assuming the bit string has the value of *010001000111011011010100100*, we break it down to sets of three bits each.

Using Eq. 2, bit code *000* corresponds to sink. Bit code *001* corresponds to Relay 1. Bit code *010* corresponds to Relay 2. Bit code *011* corresponds to Relay 3. Bit code *100* corresponds to Relay 4. Bit code *101* corresponds to Relay 5. Bit code *110* corresponds to sink again, because $(base_{10}(110)_2) \mod 6 \equiv 0$. Bit code *111* corresponds to relay-1 because $(base_{10}(111)_2) \mod 6 \equiv 1$.

Therefore according to the decoded chromosome, Sensor Node 1 and Sensor Node 7 are connected to Relay 2. Sensor Node 2 and Sensor Node 4 are connected to Relay 1. Note that although their bit codes are different, yet both bit code *001* and bit code *111* when given as input to Eq. 2 give Relay 1 as output. Sensor Node 3 is directly connected to the sink. Sensor Node 5 and Sensor Node 6 are connected to Relay 3. Sensor Node 8 and Sensor Node 9 are connected to Relay 4. There is no sensor node connected to Relay 5; therefore, it does not take part in data forwarding. The logical network is represented in Fig. 7. Note that both chromosomes of Table I and Table II are sample chromosomes chosen randomly. They do not represent the optimal solution.

The function *Translate_Chromosome()*, as shown in Algorithm 1, is responsible for chromosome translation. It takes the bit-string as input and converts it into parameters of network topology that understandable by the network simulator. In the next step, function *Network_Simulator()* calls the network simulator and gives the topology and other network parameters as input to it. The network simulator evaluates the fitness of input topology by running a network simulation and extracting results. The output of *Network_Simulator()* consists of a measure of energy remaining in the simulated network. This is called Total Residual Energy (*TRE*). Using TRE as input, the fitness of a network represented by a chromosome x is rated according to Eq. 3.

$$Fitness(x) = \min(-TRE(x)) \tag{3}$$

As *TRE* is the total remaining energy of the topology simulated by *Network_Simulator()*, a lower value of *TRE* represents lower network lifetime. A high value of *TRE* would mean that

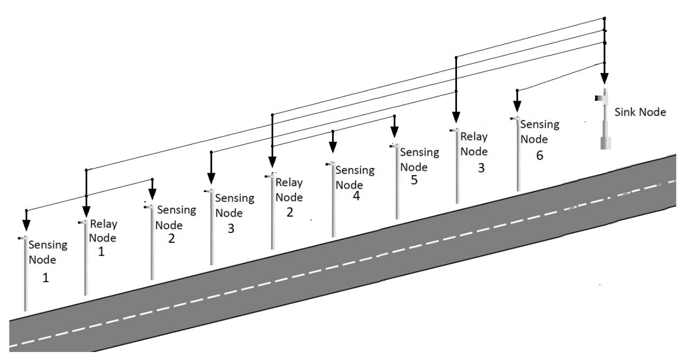


Fig. 6. Representation of LWSN with N = 6

TABLE I. DECODING A CHROMOSOME OF N = 6 NODES (10 NODES INCLUDING RELAYS AND SINK) DEPICTED IN FIG.6

Bit Position	1-2	3-4	5-6	7-8	9-10	11-12
Chromosome	01	01	11	10	10	00
Sensor Node	1	2	3	4	5	6
Decoding	Relay-1	Relay-1	Relay-3	Relay-2	Relay-2	Sink

TABLE II. DECODING A CHROMOSOME OF N = 9 NODES (15 NODES INCLUDING RELAYS AND SINK) DEPICTED IN FIG.7

Bit Position	1-3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25-27
Chromosome	010	001	000	111	011	011	010	100	100
Sensor Node	1	2	3	4	5	6	7	8	9
Decoding	Relay-2	Relay-1	Sink	Relay-1	Relay-3	Relay-3	Relay-2	Relay-4	Relay-4

the nodes of the network have more energy remaining after the simulation has run. Therefore, according to Eq. 3, higher TRE represents a lower value of fitness, which results in better fitness.

C. Genetic Operations

This subsection explains how genetic operations are carried out in OpToGen. EC requires evolutionary or genetic operators to simulate the process of evolution. Parents are selected from the current population (i.e., set of candidate solutions) for breeding. We use *Select_Random_Parent()* to select a parent for breeding in the current implementation of OpToGen. This function can be refined in future versions of OpToGen. After two parents are selected, firstly, a crossover between the parents is carried out based on the probability of crossover, P_C. The resultant offspring are subjected to mutation based on the probability of mutation, P_M. Finally, the current population is extended by adding this mutated offspring. These genetic operations are applied *I* times.

The genetic operations of crossover and mutation result in an updated population of size *2I*. These are candidates for the next generation. Now, the fitness of newly added offsprings needs to be assessed. This is done by translating the offspring chromosomes into network topologies using *Translate_Chromosome()* and then evaluated using *Network_Simulator()*. The parents of the current generation were assessed at the start of the algorithm. Both parents and offsprings are sorted based on fitness using *Sort_Decending()*. The next generation will consists of $R\%$ of stronger offsprings and *(100 - R) %* of stronger parents. This results in a new generation of size *I*.

Sorting also gives us the best (*Best_Topology*) along with its fitness (*Best_Fitness*) for each generation. The above process is carried out *G* times. *Best_Topology* either remains same or improves with each iteration.

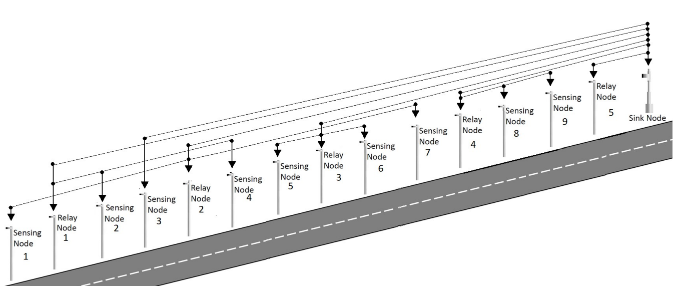


Fig. 7. Representation of LWSN with N = 9

IV. EVALUATION

To quantify the usefulness for OpToGen, we implemented OpToGen in Matlab. We used NS-2 network simulator. The OpToGen's implementation in Matlab takes the network parameters as input and generates the initial population. Then it saves the network parameters in a text file that is used with TCL script to run NS-2. The implementation is OS independent. The results of NS-2 are then written into a text file that Matlab uses for further processing.

A. Experiments with different Combinations of P_C and P_M

Matlab was used for OpToGen's performance evaluation. Best fitness corresponds to optimum network topology. For this purpose, we performed experiments with the parameter shown in Table III.

TABLE III. SIMULATIONS PARAMETERS

Parameter	Range / Value
N	6
G	30
I	20
P_C	0.6 - 0.9
P_M	0.1 - 0.5
R	30%

We ran twenty simulations for each pair of $< P_C, P_M >$ to find average fitness per $< P_C, P_M >$ pair, best fitness per $< P_C, P_M >$ pair and best fitness per generation per $< P_C, P_M >$ pair.

Fig. 8 - Fig. 10 show the results of experiments to find best fitnesses mentioned above against different values of P_C and P_M. In these figures, P_C is along the x-axis, P_M is along y-axis, and best fitness is along the z-axis. The surface contains points for all experiments.

According to Fig. 8, the first set of 20 experiments is performed with $< P_C = 0.6$, $< P_M = 0.1$ and the topology corresponding to best fitness of -4.66 was found. Note that, this is the average of best fitnesses of all 20 experiments with the same $< P_C, P_M >$ pair. Similarly, the second set of 20 experiments is performed with $< P_C = 0.6, P_M = 0.2$, and the best fitness is found to be -4.67 and so on.

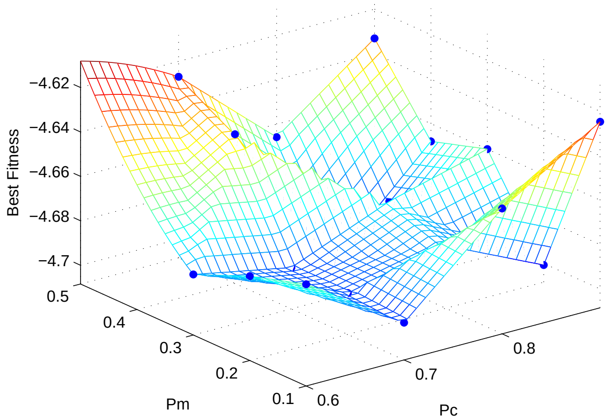


Fig. 8. Best Fitness

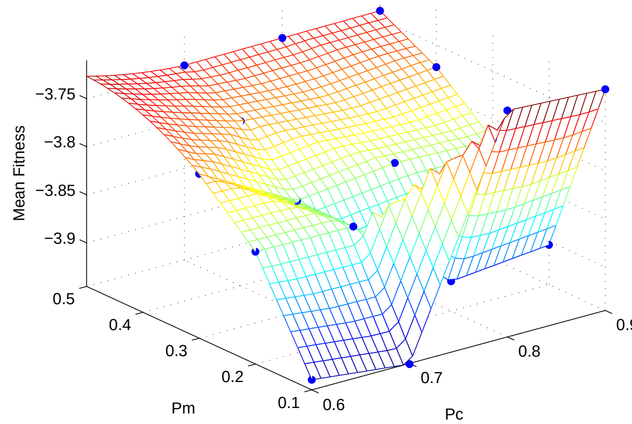


Fig. 10. Mean Fitness per Generation

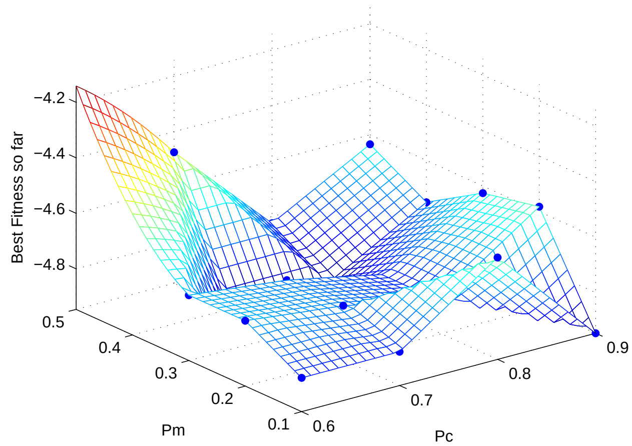


Fig. 9. Best Fitness per Generation

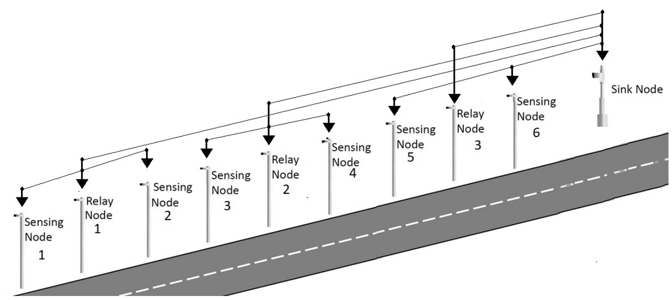


Fig. 11. Logical Representation of LWSN with best fitness

Fig. 9 indicates best fitnesses per generation for all these experiments, and Fig.10 shows mean fitness for the same. In this way, we show the convergence of OpToGen using different $< P_C, P_M >$ pairs.

During these experiments, it was seen that with the pair $< 0.6, 0.3 >$, a chromosome with the best fitness of 4.7 was created by OpToGen in the 9th generation. This means that in only *9 X 20 = 180* runs of NS-2, OpToGen converged to the best fitness.

Here, we would like to compare the resultant topology of OpToGen with a random topology from the first generation (depicted in Table I). The bit string with the best fitness had the value of *100101100000*. Table IV shows the decoded value of the chromosome, and Fig. 11 logically depicts the network with the best fitness. This can be compared to the random topology depicted in Fig. 4 and explained through Table I. The fitness of the random topology with chromosome *01111101000*, when running through NS-2, gave a value of -3.1.

B. Comparison with Brute Force

Now, we look into the probability of finding the best fitness. For this, we compared the best fitness achieved from OpToGen keeping *N = 6* for all possible combinations of networks. As chromosome's bit-string length is 12, brute force search was carried out by running NS-2 up to 2^{12} (i.e., 4096) times, covering all possible topology variations.

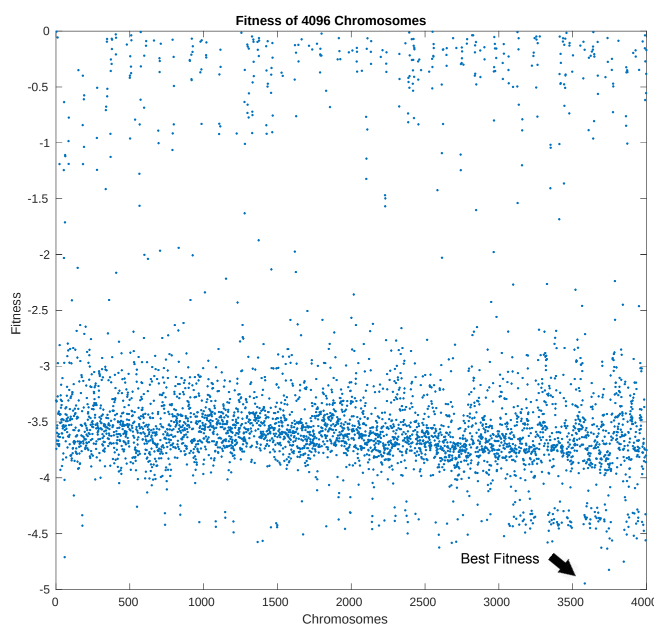


Fig. 12. Fitnesses of all possible topologies with N = 6

As depicted in Fig. 12, it took 3062 out of 4096 simulations to find out the chromosome with the best topology.

TABLE IV. CHROMOSOME REPRESENTING TOPOLOGY WITH BEST FITNESS

Bit Position	1-2	3-4	5-6	7-8	9-10	11-12
Chromosome	10	01	01	10	00	00
Sensor Node	1	2	3	4	5	6
Decoding	Relay-2	Relay-1	Relay-1	Relay-2	Sink	Sink

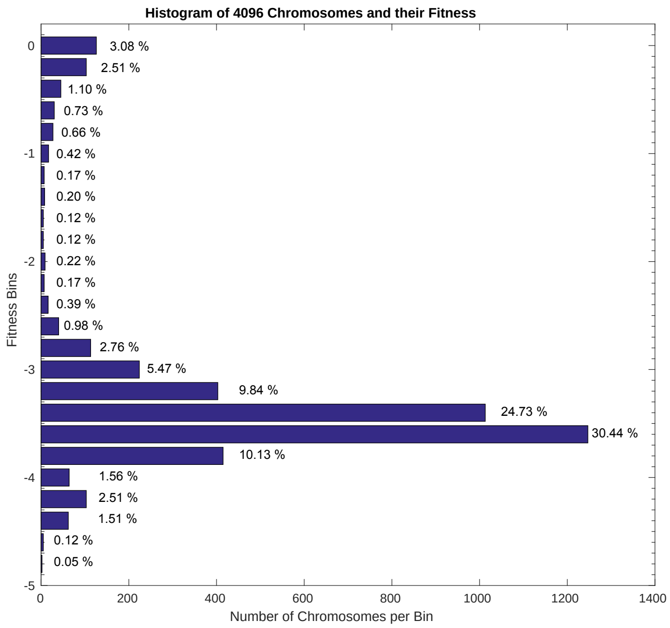


Fig. 13. Histogram of Fitnesses

A histogram of fitnesses of all 4096 simulations is shown in Fig. 13. Out of 4096 chromosomes, 3312 chromosomes had a fitness between -3 and -4, whereas, 23 chromosomes lie below -4.5. Table V shows the probability of choosing a random chromosome from all possible combinations.

TABLE V. PROBABILITY OF RANDOMLY SELECTING A CHROMOSOME WITH OPTIMAL FITNESS

Range	Count	Probability
-3 to -4	3312	80 %
-4.5 to -4.75	20	0.49 %
less than -4.75	3	0.00073 %

Hence, there was a probability of 0.00073% that a randomly selected chromosome had optimal topology. This shows how quickly OpToGen converged the best or near best as compared to brute force search.

C. Evaluation of Residual Energy

Let us compare the lifetime of OpToGen's best topology with the topologies of networks with mean fitness and least fitness, respectively. Fig. 14 shows the residual energy of three networks. The red dashed-and-dots line shows how and when the energy of the topology with the least fitness was consumed. It can be seen that the network was entirely depleted by 22-time units of NS-2 simulation. The blue dashed line shows the residual energy of the network with mean fitness. The mean fitness topology had a lifetime of 26.5-time units of NS-2 simulations. On the other hand, the best topology that OpToGen chose (depicted in Table IV earlier) lasted till 28-time units. In these cases, the topology from OpToGen ran 27% of the time more than the least fitness topology.

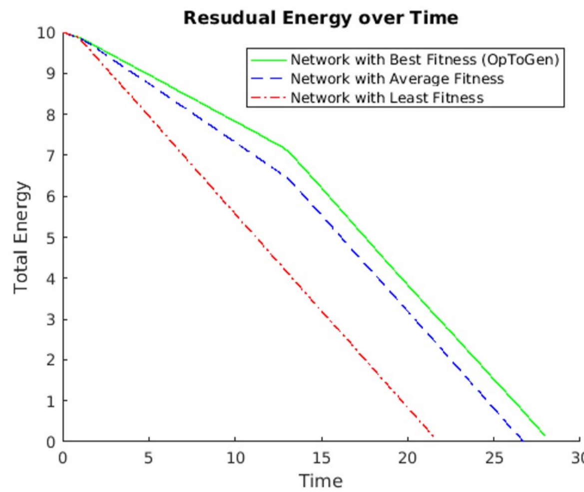


Fig. 14. Comparison of Residual Energy

TABLE VI. FITNESS CONVERGENCE VERSES NS-2 SIMULATION RUNS

Generation	P_C	P_M	Best Convergence	Simulations Required
1	0.6	0.1	18th	360
2	0.6	0.2	18th	360
3	0.6	0.3	9th	180
4	0.6	0.4	9th	180
5	0.6	0.5	9th	180
6	0.7	0.1	14th	280
7	0.7	0.2	1st	20
8	0.7	0.3	17th	340
9	0.7	0.4	4th	80
10	0.7	0.5	18th	360
11	0.8	0.1	13th	260
12	0.8	0.2	18th	360
13	0.8	0.3	8th	160
14	0.8	0.4	12th	240
15	0.8	0.5	23rd	460
16	0.9	0.1	3rd	60
17	0.9	0.2	24th	480
18	0.9	0.3	14th	280
19	0.9	0.4	10th	200
20	0.9	0.5	11th	220

D. Fitness Convergence

Table VI shows the number of NS-2 runs taken by OpToGen to achieve best fitness for all $< P_C, P_M >$ pairs. For $P_C = 0.6$ and $P_M = 0.1$ best fitness is achieved in the 18th generation. OpToGen performed 360 NS-2 runs to achieve best fitness and so on. Table VI depicts suitable $< P_C, P_M >$ for OpToGen to find best fitness.

E. Energy Consumption

Energy consumption is an important parameter for performance evaluation LWSNs. We have calculated residual energy

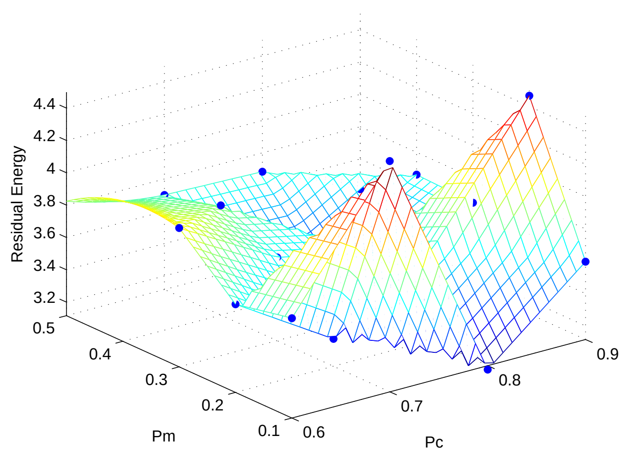


Fig. 15. Residual Energy

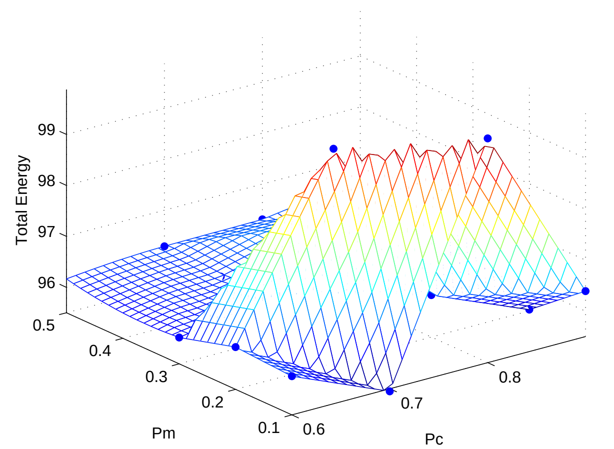


Fig. 16. Total Energy

of the LWSN, total energy consumed by the LWSN, and total energy consumed per node of LWSN for all above experiments.

1) Energy Consumption based on P_C vs P_M: Fig. 15-17 show energy consumption for experiments discussed above. For the first set of experiments with $P_C = 0.6$ and $P_M = 0.1$, residual energy for the LWSN was 3.73 units depicted in Fig. 15. Total energy of the LWSN was found to be 96.26 units corresponding to best fitness depicted in Fig. 16. Total energy per node was found to be 9.6 units as depicted in Fig. 17.

2) Energy Consumption vs Network Size: A set of experiments was performed to analyze the behavior of energy consumption for different network sizes. Very few efforts have been made in literature in this regard, especially for linear wireless sensor network. We repeated the above experiments for linear networks of 6 nodes to 12 nodes to calculate the total amount of energy consumed in these networks. Fig. 18 shows the findings of these experiments. It can be noticed that as the amount of energy consumed is high for large networks and the trend is the same throughout.

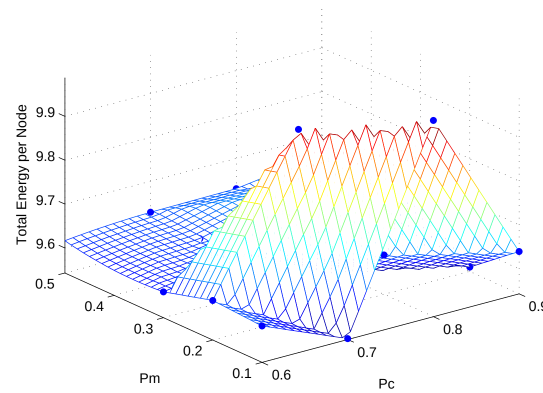


Fig. 17. Total Energy per Node

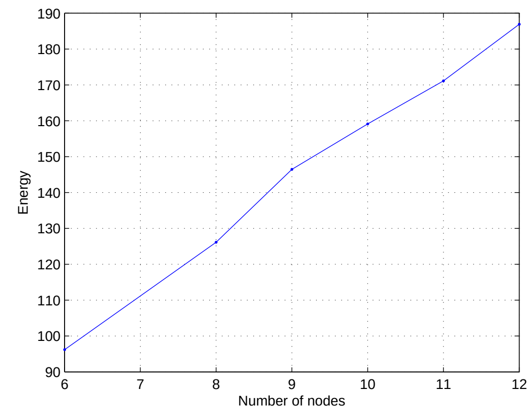


Fig. 18. Energy Consumption vs Network Size

3) Number of Generations vs Network Sizes: The final set of experiments was carried out to evaluate OpToGen's efficiency. We extracted the number of generations OpToGen took to achieve the best fitness for various network sizes. Fig. 19 shows the results of these experiments. The overall trend of the number of generations required to achieve the best fitness decreases for large networks. Therefore to achieve optimal topology in smaller networks takes the number of generations than large networks. One of the reasons for this trend is the linear nature of large networks due to which optimal topology cannot undergo much modification.

V. Conclusion

WSNs are being used to monitor structural deployments in smart cities, especially as part of intelligent transportation systems. WSNs for bridges, highways, periphery monitoring, etc. consist of a special class of WSNs called Linear Wireless Sensor Networks. Typical WSN research has focused on optimization issues related to mesh and ad hoc networks. This has left some room in the research of finding optimal topologies for LWSNs.

This paper discussed the details of OpToGen framework that can be used to find optimal topologies of LWSNs using GA. We implemented OpToGen using Matlab and NS-2. The paper also demonstrates that OpToGen proposes optimal topology in much fewer iterations than the time spent to

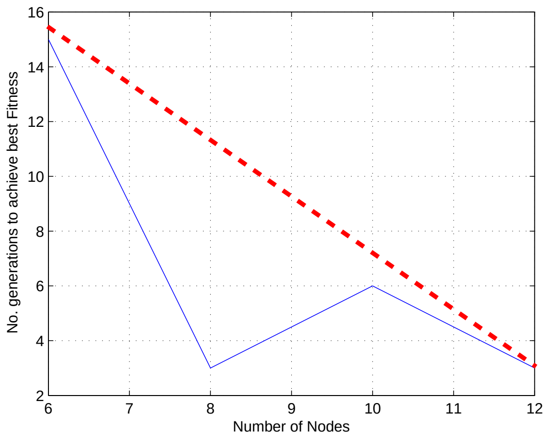


Fig. 19. Number of Generations vs Network Sizes

search the best topology, exhaustively. There are three main contributions of this paper: First, we performed experiments to make the selection of optimal probabilities easy to achieve the best fitness in minimum simulator runs. Second, the trade-off between energy consumption and different network sizes is evaluated for linear wireless sensor network. Third, we have evaluated the efficiency of OpToGen by extracting the number of generations it took to achieve the best fitness for various sized LWSNs.

VI. Future Work

There are many areas of OpToGen that require further research. These include better parent selection mechanisms and different evolutionary operators. Another area of further research is to find better ways to generate the initial population. In our current work, OpToGen's fitness function was based on network lifetime. In the future, other network parameters such as throughput, network delays, bandwidth, etc. can be made part of a multi-objective fitness function.

Acknowledgments

This work is funded by the Grant Number AT-34-335, King Abdul Aziz City of Science and Technology (KACST) of the Kingdom of Saudi Arabia. The authors thank KACST for their support.

References

[1] Zhengjie Wang, Xiaoguang Zhao and Xu Qian,: The application and issuse of linear wireless sensor networks, International Conference on System science, Engineering design and Manufacturing informatization, 2011, Guiyang, pp. 9–12

[2] Sudeep Varshney, Chiranjeev Kumar, Abhishek Swaroo, Ashish Khanna, Deepak Gupta, Joel J. P. C. Rodrigues , Placido R. Pinheiro and Victor Hugo C. de Albuquerque,: Energy Efficient Management of Pipelines in Buildings Using Linear Wireless Sensor Networks, MDPI Sensors, 2018, vol. 18, no. 8, pp. 2618

[3] Ahmed Hussein, Ahmed El–Nakib, Sherif Kishk,: Energy–efficient Linear Wireless Sensor Networks Applications in Pipelines Monitoring and Control, International Journal of Wireless Communications and Networking Technologies, 2017, vol. 6, no. 2

[4] Li R, Ma W, Huang N, Kang R,: Deployment–based lifetime optimization for linear wireless sensor networks considering both retransmission and discrete power control, PLoS ONE, 2017, vol. 12, no. 11

[5] Alfayez, F., Neill, I.,: Topology–based Optimization of Linear Wireless Sensor Networks, European Intelligence and Security Informatics Conference, 2015

[6] Mehboob, U., Qadir, J., Ali, S., Vasilakos, A.,: Genetic algorithms in wireless networking: techniques, applications, and issues, Soft Computing, 2016, vol 20, no. 6, pp 2467–2501

[7] A. A. Sheikh, E. Felemban, A. Alhindi, A. Naseer, M. Ghaleb and A. Lbath,: OpToGen – A genetic algorithm based framework for optimal topology generation for linear networks, 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, 2017, pp. 255–262.

[8] I. Jawhar, N. Mohamed, D.P. Agrawal,: Linear wireless sensor networks: Classification and applications, Journal of Network and Computer Applications, 2011, vol. 34, pp. 1671–1682

[9] Yang JB, Ke XA and Shen–Tu H.,: Optimal node placement strategies of linear wireless sensor networks with sensor node uniformly distributed, In Proceedings of the 32nd Chinese Control Conference, 2013, pp. 7365–7370

[10] Ashraf Hossain, T. Radhika, S. Chakrabarti. P. K. Biswas,: An Approach to Increase the Lifetime of a Linear Array of Wireless Sensor Nodes, International Journal of Wireless Information Networks, 2008, vol. 15, no. 2, pp. 72–81

[11] M. Alnuem,: Performance Analysis of Node Placement in Linear Wireless Sensor Networks, Journal of Emerging Trends in Computing and Information Sciences, 2014, vol. 5, no. 1

[12] Rose C.,: Low mean internodal distance network topologies and simulated annealing, IEEE Transactions on Communications, 1992, vol. 40, no. 8, pp. 1319–1326

[13] Aboeifotoh HMF and Al–Sumait LS.,: A neural approach to topological optimization of communication networks, with reliability constraints. IEEE Transactions on Reliability, 2001, vol. 50, no. 4, pp. 397–408

[14] Cheng ST.,: Topological optimization of a reliable communication network, IEEE Transactions on Reliability, 1998, vol.47, no. 3, pp. 225–233

[15] Kumar A, Pathak RM, Gupta YP,: Genetic–algorithm–based reliability optimization for computer network expansion, IEEE Transactions on Reliability, 1995, vol. 44, no. 1, pp. 63–72

[16] Innocent N, EKABUA O and Obeten,: Genetic algorithm based energy efficient optimization strategies in wireless sensor networks: A survey, Advances in Computer Science: an International Journal, 2014, vol. 3, no. 11, pp. 1–9

[17] Mojtaba Romoozi, Mehdi Vahidipour, Morteza Romoozi, Sudeh Maghsoodi,: Genetic algorithm for energy efficient and coverage–preserved positioning in wireless sensor networks, In 2010 International Conference on Intelligent Computing and Cognitive Informatics. 2010, pp. 22–25

[18] Tripathy P, Dash R and Tripathy C,: A genetic algorithm based approach for topological optimization of interconnection networks, Procedia Technology, 2012, vol 6, pp. 196–205

[19] Roslin SE.,: Genetic algorithm based cluster head optimization using topology control for hazardous environment using wsn, In 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015, pp. 1–7

[20] Kumar A, Mishra KK and Misra AK,: Optimizing the reliability of communication network using specially designed genetic algorithm, In 2009 World Congress on Nature Biologically Inspired Computing (NaBIC), 2009, pp. 499–502

[21] Clarke A and Miles J.,: Strategic fire and rescue service decision making using evolutionary algorithms, Advances in Engineering Software, 2012, vol 50, pp. 29–36

DDoS Flooding Attack Mitigation in Software Defined Networks

Safaa MAHRACH[1]
Computer, Networks, Mobility
and Modeling laboratory
Faculty of Sciences and Technology,
Hassan 1st University, Settat, Morocco

Abdelkrim HAQIQ[2]
Computer, Networks, Mobility
and Modeling laboratory
FST, Hassan 1st University, Settat, Morocco
e-NGN Research Group, Africa and Middle East

***Abstract*—Distributed denial of service (DDoS) attacks which have been completely covered by the security community, today pose a potential new menace in the software defined networks (SDN) architecture. For example, the disruption of the SDN controller could interrupt data communication in the whole SDN network. DDoS attacks can produce a great number of new and short traffic flows (e.g., a series of TCP SYN requests), which may launch spiteful flooding requests to overcharge the controller and cause flow-table overloading attacks at SDN switches. In this research work, we propose a lightweight and practical mitigation mechanism to protect SDN architecture against DDoS flooding threats and ensure a secure and efficient SDN-based networking environment. Our proposal extends the Data Plane (DP) with a classification and mitigation module to analyze the new incoming packets, classify the benign requests from the SYN flood attacks, and perform the adaptive countermeasures. The simulation results indicate that the proposed defending mechanism may efficiently tackle the DDoS flood attacks in the SDN architecture and also in the downstream servers.**

Keywords*—Software Defined Networks (SDN); Distributed Denial of Service (DDoS); network security; P4 language; DDoS mitigation

I. Introduction

Software-Defined Networking (SDN) is an emerging network architecture that enables dynamic and efficient programmability, management, and provision of the networks. This helps managers control the entire network systematically and globally, regardless of the underlying network infrastructure. SDN capabilities, such as network-visibility, centralized control, programmability, software-based traffic analysis enable the network administrator to implement easily in-network security functions, such as firewall, intrusion detection system (IDS), intrusion prevention system (IPS), etc. In this sense, many research works have developed SDN-based frameworks for enhancing network-security in smart grids, IoT, cloud, etc. [1], [2], [3], [4]. Although SDN features help in managing and protecting large networking environments against various threats, SDN itself suffers from both the present security attacks and new issues. Due to the centralized controller and the network programmability of SDN, new security challenges emerged including, malicious applications, controller-aimed distributed denial of service (DDoS) attacks, Fraudulent flow rules, flow-table overloading attacks, etc [7], [5], [6], [22].

The DDoS attacks which have been completely covered by the security community, today pose a potential new menace to the availability and scalability of the SDN network management. For example, an attacker can make the controller unavailable by producing a series of new TCP SYN requests, which involve the controller in a series of useless processes. In this case, the controller will be saturated and so unable to process legitimate requests. Moreover, an attacker may harness the limited storage capacity of the switch flow tables and launch the saturation attack (i.e. flow-table overloading attack). At the same time, SDN data plane presents a high throughput packet processing and filtering capacities which can be used to detect and mitigate DDoS attacks. In this research work, we exploit the SDN data plane capabilities to design a lightweight and practical DDoS mitigation mechanism for protecting the SDN architecture and ensure a secure and efficient SDN-based networking environment.

We extend the Data Plane (DP) with a classification and mitigation module to analyze the new incoming packets, classify the benign requests from the SYN flood attacks, and perform the adaptive countermeasures. Since the data plane has limited memory, we opted to use the stateless SYN cookie [30] technique which doesn't require storage of TCP connection states. We implement and perform the proposed mitigation prototype in P4 language [17] using the behavioral model (bmv2) software switch and Mininet emulator [16]. The simulation results indicate that the proposed defending mechanism may efficiently tackle the DDoS flood attacks in the SDN architecture and also in the downstream servers.

Benefits of the proposed approach:

- Defend the data plane from saturation attacks using SYN cookie, as a stateless technique, which doesn't require storage of TCP connection states.
- Discharged the communication path between control and data planes by performing detection, classification, and mitigation functions at the switch level.
- Protect the centralized controller and the downstream servers from the flooding attacks which affect their availability and scalability.
- Enhance the resistance of SDN technology against flooding DDoS attacks.

The present paper is organized as follows: In Section II we discuss how the SDN characteristics make it more vulnerable to DDoS attacks. In Section III, we evaluate some of the existing research works which propose security solution to detect and mitigate DDoS threats in SDN. We present the design and architecture of our approach "DDoS flooding attack mitigation in SDN" in Section IV. The simulation implementation and results are presented and evaluated in Section V. Finally, we give our conclusion and perspectives in Section VI.

II. Motivation

A. DDoS Impact in SDN

DDoS attacks which have been completely covered by the security community, today pose a potential new menace to the availability and scalability of the SDN network management [18], [7], [26], [27].

The SDN architecture is divided into three planes including, application plane, control plane, and data plane. All these planes and Application Programming Interfaces (APIs) can be targeted by the attackers to launch DDoS attacks.

- Control plane: The controller is the brain of the SDN architecture which provides central control over the network. Hence, it could be seen as a single point of failure if it is made unreachable by a Distributed Denial of Service (DDoS) attack. For example, an attacker can make the controller unavailable while producing a series of new TCP SYN requests, from distributed bot clients, which involves the controller in a series of useless processes. In this case, the controller will be saturated and so unable to process legitimate requests. In the control plane, DDoS attacks can target controller services, northbound interface, southbound interface, eastbound interface, and westbound interface.
- Data plane: The state-full implementations and functionalities in SDN like connection tracking (conntrack) [29] in Open vSwitch, OpenState [28], registers and counters in P4 [17], all require storing the state of flow in the data plane. Therefore, forwarding elements are again vulnerable to the saturation attack.
- Application plane: In the application plane, unauthenticated and unauthorized applications pose a great challenge for SDN. An attacker can launch an unauthenticated and unauthorized application with malicious programs running on the network devices and so the attacker can easily gain control of the network. Furthermore, the problem of isolation of applications and resources is not well solved in SDN; as a result, a DDoS attack on one application can affect other applications as well.

As we have seen above, DDoS attacks have a great impact on SDN application, control, and data planes. These challenges motivate us to make a study of the state-of-the-art of the detection and mitigation methods and techniques [9], [10], [11], [12], [14], [15] proposed to address DDoS impact in SDN network.

III. Related Work

Most of the research projects [23], [24], [25], [2], [13] develop defending mechanism against DDoS attacks using SDN whereas, SDN architecture itself suffer from different kind of DDoS attacks. A limited number of works [9], [10], [11], [12], [14], [15] address the potential challenges to mitigate DDoS attacks in SDN. Among which we present and analyze the following research works:

Mousavi et al. [9] exploited the SDN controller's broad network feature to protect SDN architecture from DDoS attacks. They proposed an entropy-based solution, that works within the controller, to detect DDoS attacks against SDN controllers. The proposed algorithm adds two functions to the controller; One is collecting the new incoming packets to the destination IP addresses into window size 50 and the other computes the entropy of each window and compares it to an experimental threshold. In [10] Tran et al. proposed the ODL-ANTIFLOOD solution which detects and mitigates the Controller-aimed DDoS flooding attacks. The proposed detection technique is developed based on the combination of entropy and packet in rate methods. The mitigation technique includes three steps: identify attack sources, mitigating the impact of attacks, and recovering SDN system after the attacks. Tank et al. [11] suggested an SDN-based Network Intrusion Detection System architecture . The proposed NIDS takes advantage of software-based traffic analysis and logical centralized control of SDN for detecting intrusion. They constructed a simple Deep Neural Network (DNN) for the intrusion detection system using the NSL-KDD Dataset. In [12] Li et al. applied the deep learning model to detect DDoS attacks in Software Defined Networking (SDN) environment using the ISCX data set. After the collection and analysis of network traffic feature information, the deep learning model is used for feature reduction and DDoS attack detection.

Their suggested solutions involve the controller in every operation to detect and mitigate DDoS attack, which may overload the control and data planes path and creates a potential bottleneck that impacts the network performance and restricts the scalability and reactivity of SDN controller. Moreover, the [11], [12] use complex method (i.e., deep learning) which requires a high memory and processing requirements.

Shin et al. [14] presented Avant-guard as a detection and prevention solution against the TCP SYN flooding attack. Avant-guard extends the data plane with the connection migration module which proxying the incoming TCP SYN packets and prevent the control plane from saturation attack. LineSwitch [15] is an improved Avant-guard solution. LineSwitch addressed the vulnerabilities and limitations of Avant-guard by proxying a minimum number of TCP SYN requests. The implementation of [14], [15] solutions requires complex design and several lines to add and modify to extend and customize the data-plane.

IV. System Design

Distributed Denial of Service (DDOS) is one of the former and the most common attacks which is increasing in size and frequency. In the past year, the cybersecurity vendor Akamai recorded hundreds of DDoS attacks per week. Recently, Kaspersky Labs observe an 84 percent rise in the number of DDoS attacks during the first three months of 2019 [8].

As presented in the literature, and as we discussed above in Section II, DDoS attacks are considered among the major security threats that exploit SDN vulnerabilities. The SDN

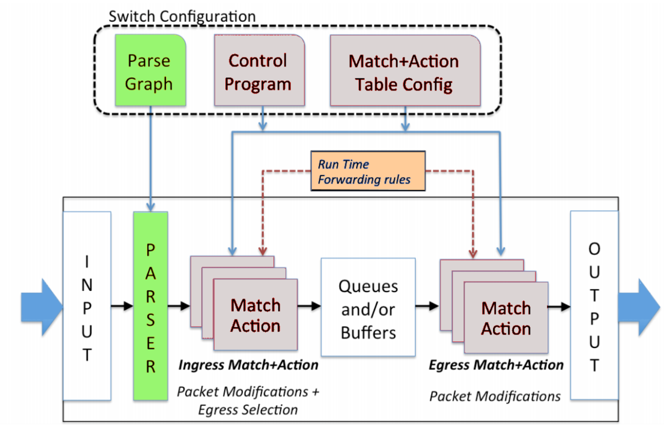


Fig. 1. Abstract forwarding model

centralized controller is an appealing target for DDoS threats. For example, an attacker can make the controller unavailable while producing a series of new TCP SYN requests, from multiple bot clients, which involves the controller in a series of useless processes. In this case, the controller will be saturated and unable to process legitimate requests. Moreover, an attacker may harness the limited storage capacity of the switch flow tables and launch the saturation attacks (i.e.,flow-table overloading attacks).

In this research work, we interest in implementing a lightweight and practical mechanism to protect the centralized SDN controller and the data plane from SYN flood attack, and achieve an efficient networking system that can resists against DDoS flooding attack.

The SDN data plane is a fast path; it has a high processing power which enables it to process a large number of packets in small time. This feature motivates us to perform a simple defensive mechanism (i.e., SYN cookie) in switch level to handle the incoming packets and classify the successful TCP connections from the failed one, which can be flooding attacks. Moreover, the data plane has limited memory, so we opted to use SYN cookie technique which doesn't require storage of TCP connections states.

Recently, Afek et al. [19] have implemented different SYN cookie methods, as anti-spoofing mechanisms, in OpenFlow 1.5 using Open vSwitch (OVS) and P4 to protect down-stream server. Their programming experience indicates that the implementation of novel applications which operate with new header fields in OpenFlow environment required complex design and several lines to add and modify compared to the programming in P4 switch. That experience confirm and validate for us the use of P4 programming language [17] and bmv2 software switch to implement our proposed SYN flood mitigation mechanism.

A. Selected Methods

1) SYN cookie technique: SYN cookie [30] prevents the memory consumption caused by the half-open SYN attacks (i.e., TCP SYN flood attacks). The system (i.e., server or switch in our case) enabling the SYN cookie technique intercepts the SYN request received from a client and sends back a SYN-ACK packet with a pre-generated cookie (i.e., the Initial Sequence Number (ISN)). The cookie or ISN is generated using some details of the initial SYN packet and cryptographic hashing function to make the decoding of the cookie more complicated (as shown in Section IV-A2).

If the mitigating system receives an ACK packet from the client (with pre-generated cookie +1), it checks the validation of the TCP sequence number (i.e., is the ACK-1) as shown in Section IV-A2. There are different SYN cookie methods that exist to interact between clients and servers, among which we define; TCP proxy, TCP reset, safe reset, HTTP redirect [19]. In this work, we select the TCP-Reset method (Fig. 2) to use in our active defensive mechanism against SYN flooding attacks.

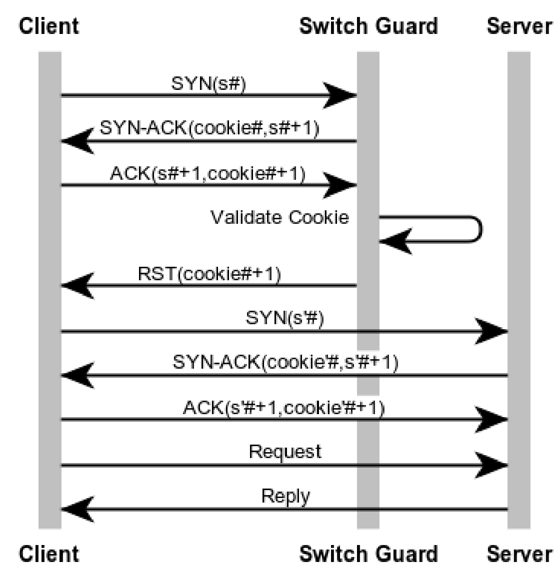


Fig. 2. TCP Reset method

TCP-reset method [31]: When the client is authenticated (i.e using SYN cookie) the mitigating system classifies it as legitimate (i.e., the system installs a rule by recording the source IP of the connection as legitimate). Then the switch sends back a TCP-reset packet (i.e., with source IP of the original server) to the client in order to enable him to re-establish the connection directly with the server. The advantage of this method is that is suitable for all TCP connections that attempt to connect when a RST packet is received.

2) Pre-generated Cookie: According to [20], the implementation of the SYN cookie technique must fulfill the following basic requirements:
- Cookies should contain some details of the initial SYN packet and its TCP options.
- Cookies should be unpredictable by attackers. It is recommended to use a cryptographic hashing function in order to make the decoding of the cookie more complicated. To this end, we select the recommended Linux SYN cookie method for generating and validating cookies [21].

TABLE I. PARAMETERS OF THE LINUX IMPLEMENTATION [21]

Parameter	Description
K_1, K_2	Secret keys
IP_s, IP_d	Source and destination IP addresses
$Port_s$, $Port_d$	Source and destination ports
ISN_s, ISN_d	Source and destination initial sequence numbers
ACK	Acknowledgement number
SEQ	Sequence number
MSS	2 bit index of the client's Maximum Segment Size
Count	32 bit minute counter
Hash()	32 bit cryptographic hash

Cookie generation:

$$H_1 = hash(K_1, IP_s, IP_d, Port_s, Port_d) \tag{1}$$

$$H_2 = hash(K_2, count, IP_s, IP_d, Port_s, Port_d) \tag{2}$$

$$ISN_d(cookie) = H_1 + ISN_s + (count \times 2^{24}) \tag{3}$$

$$+(H_2 + MSS) \mod 2^{24}$$

Cookie validation:

$$ISN_d = ACK - 1 \tag{4}$$

$$ISN_s = SEQ - 1 \tag{5}$$

$$count(cookie) = (ISN_d - H_1 - ISN_s)/2^{24} \tag{6}$$

$$MSS(cookie) = (ISN_d - H_1 - ISN_s) \tag{7}$$

$$\mod 2^{24} - H2 \mod 2^{24}$$

As we can see above and in Table I, we calculate the two hash values H1 and H2 (based on TCP options, secret keys k1, k2 and count) then we use them with ISNs and MSS to generate the cookie (ISNd), as it is shown in (3). For the cookie validation, there are 2 integrity controls (count(cookie) and MSS(cookie)). The first one checks the age of the cookie. The second evaluates whether the value of the MSS is within the 2 bit range (0-3). If the cookie meets both integrity controls, it is considered valid, and the connection can be accepted.

B. System Architecture

The goal of our proposed DDoS defensive mechanism is to activate SDN data plane with smart and advanced functions to prevent and mitigate SYN flood attacks in SDN architecture. To reach this aim, we extend the Data Plane (DP) with a classification and mitigation module to analyze and classify the new incoming packets and perform the adaptive countermeasures. In addition, we develop and implement a simple program (i.e., in python language) in the control plane (i.e., P4Runtime) to interact with P4 switch messages (packet-in).

Classification and Mitigation

The classification and mitigation module allows the Data Plane (DP) to analyze and classify the new incoming packets (i.e., which not exist in the flow table). DP will be able to classify the benign packets from the SYN flood attacks and to execute the adaptive countermeasures.

In first, DP checks if the packet exists in the flow table if so, the packet will be immediately forwarded to the destination server. Otherwise, the DP initiates a classification phase

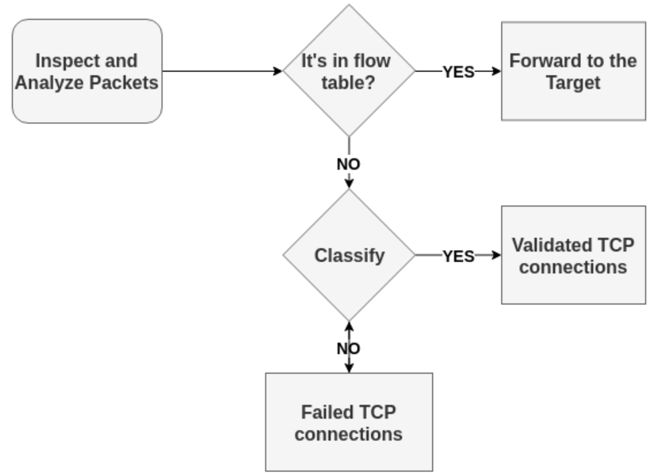


Fig. 3. TCP SYN flood defensive flowchart

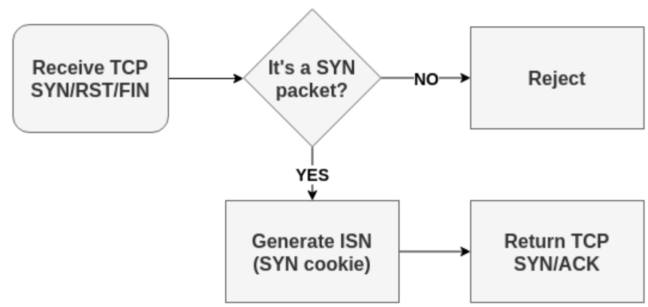


Fig. 4. Handling TCP SYN packets chart

(Fig. 3) which classifies the validated TCP connections from failed ones (i.e., half-open SYN attacks or invalidated TCP connections).

When the DP receives a TCP SYN/RST/FIN packet, as shown in Fig. 4, it checks whether it is a SYN packet. If so, the DP responds to the SYN request client by a SYN-ACK packet with a pre-generated cookie (see IV-A2). The cookie or the Initial Sequence Number (ISN) is created by hashing details about the initial SYN packet and its TCP options. If the packet is not a TCP SYN packet (i.e., TCP FIN or TCP RST), it is rejected.

If the DP receives an ACK packet, as shown in Fig. 5, it checks the validation of the TCP sequence number (i.e., is the ACK-1) as shown in Section IV-A2. For the validation of the cookie, we control two integrities; count(cookie) for checking the age of the cookie (i.e., must be lower to 2min), and MSS(cookie) for evaluating whether the value of the MSS is within the 2-bit range (0-3).

If the TCP connection is validated the switch transmits the ACK packet to the controller to write the required flow-entry. At the same time, the switch sends back a TCP-reset packet (i.e., with source IP of the original server) to the client in order to enable him to re-establish the connection directly with the server. In our system, we use the P4Runtime API as a control plane specification for controlling the data plane

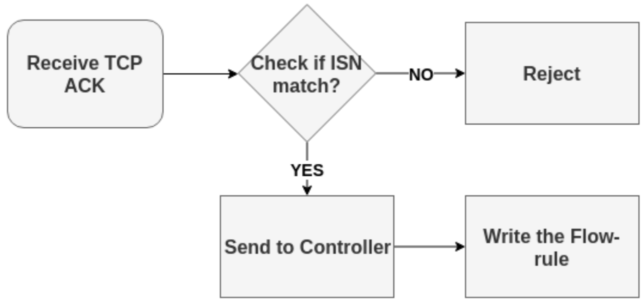


Fig. 5. Handling TCP ACK packets chart

elements executing the P4 programs.

P4Runtime API program

We develop a simple program in python language to interact with P4 switch messages (packet-in). The program connects the P4Runtime controller to the switch(s), installs the P4 program on the switch, populates the defined match action tables, and describes how the controller treats the packet-in messages.

When the controller receives a packet-in message it learns from the details of the initial SYN packet to create the requested flow-rule. It extracts the addresses of IP source, IP destination, Mac source, and Mac destination, and the ports of source and destination. It checks if the new flow has already a flow-rule in the memory of the controller, if it exists it writes it on the switch, if not it creates a new flow rule for the new flow and writes it on switch flow-table. Each flow rule is installed with a timeout value, both hard and idle timeouts, to avert the growth of rules in the switch flow-tables.

V. Simulation

A. Environment

We use the Behavioral model-bmv2 framework which supports the P4 software switch, with v1model architecture. This reference implementation covers P4.16 specification version 1.2.0 [17], and it functions as the data plane. We exploit bmv2 for using two targets `simple_switch` (i.e., this target is a software switch, running on a general-purpose CPU, such as Intel/AMD/etc., that can execute `P4_14` and `P4_16` programs) and `simple_switch_grpc`. The difference between both of the targets is that `simple_switch_grpc` can also accept TCP connections from a controller, where the format of control messages on this connection is defined by the P4Runtime API specification. The P4Runtime API is a control plane specification for controlling the data plane elements of a device or program defined by a P4 program. P4c is used as a compiler; it compiles the P4 program (.p4) into the JSON file to be implemented on the P4 software switch, and defines the tables and actions, in Protobuf format(), existing in P4 program to be populated by the controller.

B. Implementation

Our focus here is the software-based implementation, used to direct the packet handling process inside the switch, unlike the traditional SDN data plane which requires hard and heavy changes to extend and customize the data-plane. Our approach is based upon the Abstract forwarding model of P4 (see Fig. 1) which consists of headers, describe the fields and their sizes of each header within a packet; parser, defines the permitted header sequences within packets; control flow, organizes the layout of match action tables within ingress and egress pipeline, and the packet flow through the pipeline; match action tables, associate user-defined keys with actions; stateful memories, counters, meters, and registers are used to store information across packets.

We develop a P4 program which describes how packets are analyzed and classified benign packets from flooding attacks, and which defines the adaptive actions and countermeasures to perform. The proposed processing pipeline in the data-plane is presented below. The parser is changed to extract TCP fields and TCP options. We add three match action tables: the first one, named `ipv4_lpm()`; which handles the incoming packets and forward to the target those matching the existing flow rules, the second one, named `return_SYN_ACK()`; responds to the SYN packet (which doesn't correspond to any flow entry) by sending a `SYN_ACK` packet with a pre-generated sequence number (see Section IV-A2), the third one, named `ACK_verify()`; it checks and verify the validity of the sequence number existing in the received ACK packet. If the sequence number of the ACK packet is validated the action `send_to_cpu()` is performed to forward the validated ACK packet to the control plane for adding the requested flow entry. The control plane (P4Runtime) learns from the received ACK packet to build the requested flow entry, and then write it on the data plane for handling the incoming packets from the same flow. Optionally, the switch sends back a TCP-reset packet (i.e., with source IP of the original server) to the client to enable him to re-establish the connection directly with the server.

```
 apply {
ipv4_lpm.apply();

if (hdr.tcp.flags == 02) {
   return_synack();
}
else if (hdr.tcp.flags == 10) {
    ACK_verify();
    if ((meta.meta.count_cookie <= 2) &&
    (0 <= meta.meta.mss_cookie) &&
    (meta.meta.mss_cookie <= 3)){
send_to_cpu();}
else{
drop(); }
}}
```

C. Use Case

The test environment of this experiment includes a p4 software switch in which our `p4_program` is implemented, P4Runtime API used to populate the existing match tables, a server that hosts HTTP service, benign clients which send HTTP requests, and an attacker who performs a TCP SYN flood attack using hping3. Fig. 6 shows a new client which sent an HTTP request targeted to the web server (10.0.1.1). Since the client is new (i.e., the proper rule doesn't exist in

No.	Time	Source	Destination	Protocol	Leng	Info
1	0.00…	10.0.3.3	10.0.1.1	TCP	74	59714 → 80 [SYN] Seq=4033130284 Win=29200 Len
2	0.01…	10.0.1.1	10.0.3.3	TCP	74	80 → 59714 [SYN, ACK] Seq=6666666 Ack=4033130
3	0.01…	10.0.3.3	10.0.1.1	TCP	66	59714 → 80 [ACK] Seq=4033130285 Ack=6666667 W
4	0.01…	10.0.3.3	10.0.1.1	HTTP	201	GET / HTTP/1.1
5	0.24…	10.0.3.3	10.0.1.1	TCP	201	[TCP Retransmission] 59714 → 80 [PSH, ACK] Se
6	0.47…	10.0.3.3	10.0.1.1	TCP	201	[TCP Retransmission] 59714 → 80 [PSH, ACK] Se
7	0.92…	10.0.3.3	10.0.1.1	TCP	201	[TCP Retransmission] 59714 → 80 [PSH, ACK] Se
14	42.1…	10.0.3.3	10.0.1.1	TCP	74	59716 → 80 [SYN] Seq=1149509130 Win=29200 Len
15	42.1…	10.0.1.1	10.0.3.3	TCP	74	80 → 59716 [SYN, ACK] Seq=881920156 Ack=11495
16	42.1…	10.0.3.3	10.0.1.1	TCP	66	59716 → 80 [ACK] Seq=1149509131 Ack=881920157
17	42.1…	10.0.3.3	10.0.1.1	HTTP	201	GET / HTTP/1.1
18	42.1…	10.0.1.1	10.0.3.3	TCP	66	80 → 59716 [ACK] Seq=881920157 Ack=1149509266
19	42.1…	10.0.1.1	10.0.3.3	TCP	83	80 → 59716 [PSH, ACK] Seq=881920157 Ack=11495
20	42.1…	10.0.3.3	10.0.1.1	TCP	66	59716 → 80 [ACK] Seq=1149509266 Ack=881920174
21	42.1…	10.0.1.1	10.0.3.3	HTTP	817	HTTP/1.0 200 OK (text/html)
22	42.1…	10.0.3.3	10.0.1.1	TCP	66	59716 → 80 [ACK] Seq=1149509266 Ack=881920926
23	42.1…	10.0.3.3	10.0.1.1	TCP	66	59716 → 80 [FIN, ACK] Seq=1149509266 Ack=8819
24	42.2…	10.0.1.1	10.0.3.3	TCP	66	80 → 59716 [ACK] Seq=881920926 Ack=1149509267

Fig. 6. Classification of successful TCP connection from flood attack: How the mitigation switch traits the legitimate packets

Topic / Item	Count	Ave	Min	Max	Rate (ms)	Percent
Total HTTP Packets	34				0.0006	100%
Other HTTP Packets	0				0.0000	0.00%
HTTP Response Packets	17				0.0003	50.00%
HTTP Request Packets	17				0.0003	50.00%

Fig. 7. Wireshark statistics: HTTP packet counter in the Web server

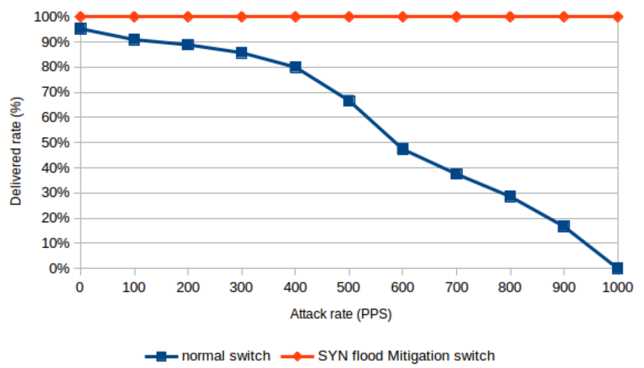


Fig. 8. Percentage of successfully delivered packets to the HTTP server from benign clients

No.	Time	Source	Destination	Protocol	Leng	Info
47097	12…	201.63.214.0	10.0.1.1	TCP	54	48046 → 80 [SYN] Seq=1124822665 Win=512 Len=0
47098	12…	55.235.228.227	10.0.1.1	TCP	54	48047 → 80 [SYN] Seq=1620906196 Win=512 Len=0
47099	12…	14.188.151.117	10.0.1.1	TCP	54	48048 → 80 [SYN] Seq=557302854 Win=512 Len=0
47100	12…	97.40.170.151	10.0.1.1	TCP	54	48049 → 80 [SYN] Seq=602198780 Win=512 Len=0
47101	12…	214.148.228.2…	10.0.1.1	TCP	54	48050 → 80 [SYN] Seq=747683763 Win=512 Len=0
47102	12…	151.154.227.1…	10.0.1.1	TCP	54	48051 → 80 [SYN] Seq=1935660469 Win=512 Len=0
47103	12…	219.7.165.148	10.0.1.1	TCP	54	48052 → 80 [SYN] Seq=1857775728 Win=512 Len=0
47104	12…	10.0.1.1	214.62.71…	TCP	54	80 → 16991 [SYN, ACK] Seq=3173826 Ack=1764704849 Win=512
47105	12…	10.0.1.1	170.84.22…	TCP	54	80 → 16992 [SYN, ACK] Seq=4838241 Ack=608783951 Win=512
47106	12…	10.0.1.1	141.113.1…	TCP	54	80 → 16993 [SYN, ACK] Seq=15222798 Ack=166182527 Win=512
47107	12…	10.0.1.1	84.109.70…	TCP	54	80 → 16994 [SYN, ACK] Seq=1181623 Ack=1611697434 Win=512
47108	12…	205.171.125.1…	10.0.1.1	TCP	54	48053 → 80 [SYN] Seq=852388624 Win=512 Len=0
47109	12…	63.177.117.113	10.0.1.1	TCP	54	48054 → 80 [SYN] Seq=582093657 Win=512 Len=0
47110	12…	179.36.113.188	10.0.1.1	TCP	54	48055 → 80 [SYN] Seq=1916942219 Win=512 Len=0

Fig. 9. How our mitigation system reacts to spoofed SYN flood attack

flow-table), the mitigation switch will respond by a SYN-ACK packet and then check the validation of the received ACK packet (see the three first lines). This latter is sent to the controller to add the appropriate flow-entry. Then, the legitimate client (10.0.3.3) connects directly to the web server and get the requested data as we can see in the lines (14-24).

In this scenario, we measure the delivered packet rate of benign clients under network saturation attack. Fig. 9 indicates a spoofed SYN flood attack targeted the web server. In this case, the mitigation switch intercepts and responds to the received SYN requests to validate the successful TCP connections and ignore the failed one. Even though the network is under several flood attack, the web server is available (i.e., because it receives only the authenticated TCP connections) and can process and respond to the HTTP requests from legitimate clients as shown in Fig. 7. To further show the impact of saturation attacks on normal traffic in detail, we vary the packet sending rate of the TCP SYN flood attack from 0 to 1000 per second, and at the same time, we send the HTTP requests from 15 benign clients. The test results are shown in Fig. 8.

With SYN flood mitigation switch: The SYN flood packets are blocked by the mitigation switch and so, only the validated TCP connections are transmitted to the controller for writing the new flow-rules. In this case, web server will receive only the SYN requests from benign clients. As shown in 8, the packet rate of benign clients are `100%` delivered to the web server, even while the network is under a severe network saturation attack.

With normal switch; All the SYN requests, from the attacker, are forwarded to the controller as packet-in messages and so, the controller is saturated and the flow table is also overloaded, which explain, the packet rate of benign clients are nearly `0%` delivered.

D. Evaluation and Comparison

Under SYN flooding attack the data plane receives many TCP SYN packets, from random IP sources, that it will then transmit to the controller. These TCP requests may saturate the channel between control and data planes, the couple of SDN controller and switches, and also the downstream servers.

The proposed SYN flood mitigation mechanism enables the data plane to handle the new incoming packets and forward to the controller only the requests of the successful TCP connections. It discharges the controller-switch path and reduces the involvement of SDN controller. Thus, the SDN network will be more protected from Controller-aimed distributed denial of service (DDoS) attack and flow-table overloading attack. Moreover, it defends the downstream servers from DDoS flooding attacks by replying unconditionally to the received TCP SYN requests. Consequently, the proposed mechanism makes the SDN network more scalable and efficient to resist such attacks.

Our prototype uses the SYN cookie technique which doesn't require the storage of network states in the data plane. This is in contrast to previous works, such as Avant-guard [14], which use SYN proxy method to protect SDN controller against SYN flood attacks. The SYN proxy requires the information storage (timestamp, sequence number, source IP and port) throughout the TCP connections, which gives rise to a new type of SYN flooding attack named Buffer Saturation Attack.

The implementation of the proposed method uses software-based environment to direct the packet handling process inside the switch, unlike the traditional hardware-based implementation which requires hard and heavy changes to extend and customize the data-plane.

In comparison to some existing solutions, our approach implements simple method functionalities over SDN data plane rather than complex methods, such as machine learning or deep learning which require high memory and processing requirements [11], [12].

Moreover, it enhances the resilience against TCP SYN Flood attack and it may also be used to defence attacks based on HTTP, SMTP, and other protocols. This can be done while using specific SYN cookie methods, such as HTTP Redirect and TCP Safe Reset [19].

VI. Conclusion

The growing adoption of SDN technology in controlling and securing the legacy networks, such as cloud, IOT, cyber-physical systems have highlighted the requirement to analyze and evaluate the benefits and vulnerabilities of SDN architecture. While DDoS attacks remain a top threat that is growing in size and frequency of reported incidents, SDN opens the door for yet new vulnerabilities to this type of attacks. In this paper, we have examined how the SDN characteristics make it more vulnerable to DDoS attacks and we have evaluated some of the existing research works which propose solutions to detect and mitigate DDoS threats in SDN. In this sense, we have presented and implemented a lightweight and practical mitigation mechanism to protect SDN network against DDoS flooding attack. The proposed approach enables the data plane to analyze the new incoming packets, classify the benign requests from the SYN flood attacks, and perform the adaptive countermeasures.

In comparison to the existing solutions, our approach activates the mitigation of DDoS flooding attack at the SDN data plane without any external and dedicated appliance. Consequently, it prevents and reduces the risk of saturation attack in the SDN controller and switches. Moreover, the simulation results indicate that the proposed mechanism may efficiently tackle the DDoS flood attacks in both SDN architecture and downstream servers. For future work, we plan to conduct a study of the different network systems, such as cloud and IoT to decide which one to choose as an application domain in our experiments. Moreover, further experiments and simulations will be performed to support more sophisticated attacks.

References

[1] Maziku, H., Shetty, S. and Nicol, D.M., 2019. Security risk assessment for SDN-enabled smart grids. Computer Communications, 133, pp.1-11.

[2] Mahrach, S., El Mir, I., Haqiq, A. and Huang, D., 2018. SDN-based SYN flooding defense in cloud. Journal of Information Assurance and Security, 13(1).

[3] Gonzalez, C., Charfadine, S.M., Flauzac, O. and Nolot, F., 2016, July. SDN-based security framework for the IoT in distributed grid. In 2016 International Multidisciplinary Conference on Computer and Energy Science (SpliTech) (pp. 1-5). IEEE.

[4] Zaalouk, A., Khondoker, R., Marx, R. and Bayarou, K.M., 2014, May. OrchSec: An orchestrator-based architecture for enhancing network-security using Network Monitoring and SDN Control functions. In NOMS (pp. 1-9).

[5] Zhang, H., Cai, Z., Liu, Q., Xiao, Q., Li, Y. and Cheang, C.F., 2018. A survey on security-aware measurement in SDN. Security and Communication Networks, 2018.

[6] Rawat, D.B. and Reddy, S.R., 2016. Software defined networking architecture, security and energy efficiency: A survey. IEEE Communications Surveys & Tutorials, 19(1), pp.325-346.

[7] Mahrach, S. and Haqiq, A., 2019. DDoS Defense in SDn-Based Cyber-Physical Cloud. Cybersecurity and Privacy in Cyber Physical Systems, p.133.

[8] DDoS Attacks in Q1 2019 report. https://securelist.com/ddos-report-q1-2019/90792/.

[9] Mousavi, S.M. and St-Hilaire, M., 2018. Early detection of ddos attacks against software defined network controllers. Journal of Network and Systems Management, 26(3), pp.573-591.

[10] Tran, N.T., Le, T.L. and Tran, M.A.T., 2018, November. ODL-ANTIFLOOD: A Comprehensive Solution For Securing OpenDayLight Controller. In 2018 International Conference on Advanced Computing and Applications (ACOMP) (pp. 14-21). IEEE.

[11] Tang, T.A., Mhamdi, L., McLernon, D., Zaidi, S.A.R. and Ghogho, M., 2016, October. Deep learning approach for network intrusion detection in software defined networking. In 2016 International Conference on Wireless Networks and Mobile Communications (WINCOM) (pp. 258-263). IEEE.

[12] Li, C., Wu, Y., Yuan, X., Sun, Z., Wang, W., Li, X. and Gong, L., 2018. Detection and defense of DDoS attack–based on deep learning in OpenFlow-based SDN. International Journal of Communication Systems, 31(5), p.e3497.

[13] Mahrach, S., Mjihil, O. and Haqiq, A., 2017, December. Scalable and Dynamic Network Intrusion Detection and Prevention System. In International Conference on Innovations in Bio-Inspired Computing and Applications (pp. 318-328). Springer, Cham.

[14] Shin, S., Yegneswaran, V., Porras, P. and Gu, G., 2013, November. Avant-guard: Scalable and vigilant switch flow management in software-defined networks. In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security (pp. 413-424). ACM.

[15] Ambrosin, M., Conti, M., De Gaspari, F. and Poovendran, R., 2015, April. Lineswitch: Efficiently managing switch flow in software-defined networking while effectively tackling dos attacks. In Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security (pp. 639-644). ACM.

[16] Mininet emulator. http://mininet.org/.

[17] The P4 Language Specification. https://p4.org/p4-spec/docs/P4-16-v1.2.0.pdf.

[18] Yan, Q., Yu, F.R., Gong, Q. and Li, J., 2015. Software-defined networking (SDN) and distributed denial of service (DDoS) attacks in cloud computing environments: A survey, some research issues, and challenges. IEEE Communications Surveys & Tutorials, 18(1), pp.602-622.

[19] Afek, Y., Bremler-Barr, A. and Shafir, L., 2017, May. Network anti-spoofing with SDN data plane. In IEEE INFOCOM 2017-IEEE Conference on Computer Communications (pp. 1-9). IEEE.

[20] Simpson, W.A., 2011. TCP cookie transactions (TCPCT).

[21] Echevarria, J.J., Garaizar, P. and Legarda, J., 2018. An experimental study on the applicability of SYN cookies to networked constrained devices. Software: Practice and Experience, 48(3), pp.740-749.

[22] Zhang, H., Cai, Z., Liu, Q., Xiao, Q., Li, Y. and Cheang, C.F., 2018. A survey on security-aware measurement in SDN. Security and Communication Networks, 2018.

[23] Swami, R., Dave, M. and Ranga, V., 2019. Software-defined Networking-based DDoS Defense Mechanisms. ACM Computing Surveys (CSUR), 52(2), p.28.

[24] Birkinshaw, C., Rouka, E. and Vassilakis, V.G., 2019. Implementing an intrusion detection and prevention system using software-defined networking: Defending against port-scanning and denial-of-service attacks. Journal of Network and Computer Applications, 136, pp.71-85.

[25] Rebecchi, F., Boite, J., Nardin, P.A., Bouet, M. and Conan, V., 2019. DDoS protection with stateful software-defined networking. International Journal of Network Management, 29(1), p.e2042.

[26] Dong, S., Abbas, K. and Jain, R., 2019. A survey on distributed denial of service (DDoS) attacks in SDN and cloud computing environments. IEEE Access, 7, pp.80813-80828.

[27] Krishnan, S. and Oliver, J.J.E., 2019, April. Mitigating DDoS Attacks in Software Defined Networks. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 960-963). IEEE.

[28] Bianchi, G., Bonola, M., Capone, A. and Cascone, C., 2014. OpenState: programming platform-independent stateful openflow applications inside the switch. ACM SIGCOMM Computer Communication Review, 44(2), pp.44-51.

[29] The connection tracking system. http://docs.openvswitch.org/en///latest/tutorials/ovs-conntrack/

[30] Fontes, S.M., Hind, J.R., Narten, T. and Stockton, M.L., International Business Machines Corp, 2006. Blended SYN cookies. U.S. Patent 7,058,718.

[31] Touitou, D., Pazi, G., Shtein, Y. and Tzadikario, R., Cisco Technology Inc, 2011. Using TCP to authenticate IP source addresses. U.S. Patent 7,979,694.

AUTHORS' PROFILE

Safaa Mahrach has received Master degree in Networks and Computer Systems in 2013 from Hassan 1^{st} University of Settat, Morocco, and she is currently working toward the PhD degree from Computer, Networks, Mobility and Modeling laboratory, Mathematic and Computing Department at Hassan 1^{st} University, Morocco. His research interests are Security of computer networks and Cloud systems using the Software Defined Networking (SDN) paradigm and the P4 programming language. She is member of the P4 Language Consortium.

Abdelkrim HAQIQ has a High Study Degree and a PhD , both in the field of modeling and performance evaluation of computer communication networks, from the University of Mohammed V, Agdal, Faculty of Sciences, Rabat, Morocco. Since September 1995 he has been working as a Professor at the department of Mathematics and Computer at the Faculty of Sciences and Techniques, Settat, Morocco. He is the Director of Computer, Networks, Mobility and Modeling laboratory. He is also the General Secretary of the electronic Next Generation Networks (e-NGN) Research Group, Moroccan section. He is an IEEE Senior member and an IEEE Communications Society member. He was a co-director of a NATO Multi-Year project entitled "Cyber Security Analysis and Assurance using Cloud-Based Security Measurement system", having the code: SPS-984425. Dr. Abdelkrim HAQIQ's interests lie in the areas of modeling and performance evaluation of communication networks, mobile communications networks, cloud computing and security, queueing theory and game theory. He is the author and co-author of more than 150 papers (international journals and conferences/workshops). He is also a member of the board of the International Journal of Intelligent Engineering Informatics.

Investigation of Deep Learning-based Techniques for Load Disaggregation, Low-Frequency Approach

Abdolmaged Alkhulaifi[1]
Department of Information and Computer Science
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

Abdulah J. Aljohani[2]
Center of Excellence in Intelligent Engineering Systems
Department of Electrical and Computer Engineering
King Abdulaziz University, Jeddah 21589, Saudi Arabia

***Abstract*—Unlike sub-metering, which requires individual appliances to be equipped with their own meters, non-intrusive load monitoring (NILM) use algorithms to discover appliance individual consumption from the aggregated overall energy reading. Approaches that uses low frequency sampled data are more applicable in a real world smart meters that has typical sampling capability of $\leq 1Hz$. In this paper, a systematic literature review on deep-learning-based approaches for NILM problem is conducted, aiming to analyse the four key aspects pertaining to deep learning adoption. This includes deep learning model adoption, features selection that are used to train the model, used data set and model accuracy. In our study, analyses the performance of four different deep learning approaches, namely, denoising autoencoder (DAE), recurrent long short-term memory (LSTM) , Recurrent gate recurrent unit (GRU), and sequence to point. Our experiments will be conducted using the two data sets, namely, REDD and UK-DALE. According to our analysis, the sequence to point model has achieved the best results with an average mean absolute error (MAE) of 14.98 watt when compared to other counterpart algorithms.**

***Keywords*—NILM; deep learning; load disaggregation; recurrent long short-term memory; gate recurrent unit**

I. Introduction

Energy disaggregation, also called non-intrusive load monitoring NILM, is the method of decomposing the aggregated energy consumption of the whole household down to individual appliance usage. The problem was firstly introduced in Hart's seminal paper in 1992 [1], and has been investigated intensively since then. NILM aims to allow the household occupants to understand the consumption of each appliance and hence take effective action towards reducing the power consumption. Reporting individual appliance consumption could lead to energy consumption reduction by more than 15% [2], [3]. Numerous NILM algorithms have been proposed, where they can be divided into two main categories according to type of data they employ, namely: low frequency and high frequency. The former approach uses data that are collected in low sampling rate typically ($< 1Hz$), while the latter relies on data that are collected at high sampling rate ($> 50Hz$). Researchers have been focusing on low frequency approaches as they can be readily applied to current smart meters [4]–[7].

One of the most widely used metric in measuring an energy disaggregation algorithm is the mean absolute error (MAE), which can be formulates as:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|\widehat{y}_t^{(i)} - y_t^{(i)}\right| \tag{1}$$

where $y_t^{(i)}$ and $\widehat{y}_t^{(i)}$ are the actual and estimated power consumption of the i^{th} appliance at t instance, respectively. The appliances that are usually picked for testing are: kettle, microwave, fridge, dish washer and washing machine. Note that, the analysis uses MAE as one of the metric measurement.

Recently, deep learning techniques have been widely used in solving the low-frequency-based NILM problem, due to their capabilities of extracting features and patterns [4]–[9]. For example, three models were proposed in [5]: first model was based on denoising autoencoder (DAE) that is aiming to reconstruct a clean target from the noisy data input. The second was based a convolutional neural network (CNN)-trained model with aim to estimate the start time, end time and mean power demand. While the third was based on the long short-term memory (LSTM) recurrent neural network (RNN) architecture. The study has concluded that the DAE, CNN,and LSTM-based RNN architectures performed adequately well achieving MAE score of 18, 14 and 70 in watts respectively, when compared to non-deep learning-based techniques counterparts of combinatorial optimization (CO) and factorial hidden Markov model (FHMM) both achieved higher error, i.e. MAE of 70 and 170 respectively [5]. Note that, all of the discussed approaches in [5] were compared using the Domestic Appliance-Level Electricity (UK-DALE) data set [10] and using active power as input features.

II. Background

In [11], a hybrid model based on both hidden markov model (HMM) and deep neural network (DNN) was proposed. It works by training HMM with two emission probabilities, one for the single load to be extracted and the other for the aggregate power signal. To elaborate a little, Gaussian distribution was used to model observations of the single load whereas observations of the aggregate signal are modeled with a DNN. Aiming to learn more features, MoWan He et al. [12] modified the RNN of [5] by adding multiple parallel convolutional layers with varying filter size to detect features from aggregated signal. This idea was borrowed from GoogleLeNet [13] model for image recognition and it's also used in natural language processing.

All approaches so far tackled NILM as a sequence to sequence, given a sequence of aggregated power try to find the sequence of the appliance disaggregated power. However, in [6] a sequence to point model was proposed, where given a sequence of aggregated power find the mid-point in the appliance disaggregated power sequence. By applying sliding window on the aggregated data, the model will cover all points in the disaggregated signal. This new approach was compared to the autoencoder approach of [5] and has achieved a significant low error of MAE= 15.47 across all appliances compared to 93.49 achieved by DAE counterpart.

Typically, the active power which is the actual power that is consumed measured in watts, was the only feature that was used in energy disaggregation in low frequency deep learning-based approaches. However, M. Valenti et al. [7] introduced the idea of using reactive power, The wasted power resulting from inductive and capacitive loads measured in volt-amperes reactive, with the active power. Two different data-sets were used, namely UK-DALE [10] and Almanac of Minutely Power data set (AMPds) [14], where the model of in [7] was able to outperform the model proposed in [5] by around 8.4% and 8.4% using UK-DALE data set and AMPds, respectively.

D. Murray et al. [4] presented a study on the transferability of neural network approaches across different data-set. The purpose of the study was to measure the scalability of neural network approaches in large scale smart meter deployment. Two architecture were proposed, a CNN architecture with 28,696,641 parameters and a gate recurrent unit (GRU) architecture with 4,861 parameters [4]. Evaluation was conducted across three data sets: REDD data set [15], UK-DALE [10], and REFIT [16], where models were trained on one data set and tested on another. Results from [4] showed that the two proposed architecture preformed well in transferability test with minimal performance drop compared to training and testing on the same data set. Although Both GRU-based network and CNN-based network showed similar performance, the GRU-based network was easier to train and less complex due to having less trainable parameters compared to CNN. In [17], C. Shin et al. explored a new direction for energy disaggregation by combining regression and classification network. By multiplying regression output with classification probability to form the final estimates, their proposed model which is employing subtask gated networks (SGN), outputs the power estimation gated with on/off classification. In their experiment in REDD and UK-DALE data-sets, they reported that SGN showed $15 and 30\%$ improved performance on average when compared to of the FHMM [18], and DAE of [5].

Against this background, we will analyze the performance of four different deep learning approaches, namely, DAE, Recurrent LSTM, Recurrent GRU, and Sequence to point, aiming to evaluate their accuracy within the NILM problem context. Our experiments will conducted using the two, well-known, data-sets [15] and UK-DALE [10]. The rest of the paper is organised as follows. The experiment design will be detailed in Section III, in which the data sets selection criteria will be explained. In Section IV the experiment performance is quantified. Finally, our conclusions will be offered in Section V.

III. Experimental Design

This section will discuss our experiment set-up in which we selected the two most widely used data set, namely REDD [15], and UK-DALE [10]. During our experiment, we will conduct a transfer-ability test on each of the following four models:

- Denoising Autoencoder: Denoising autoencoder (DAE) was introduced by J. Kelly et al. [5]. It's a sequence to sequence model that works by attempting to reconstruct a clean target from a noisy input. They showed that denoising autoencoders performed better than other architectures for sequence to sequence learning. The Keras implementation of the model (the building of layers) was taken from a reimplementation of Kelly DAE model in Keras by Taiwan Power Company[1].
- Recurrent Neural Network (LSTM): The recurrent LSTM model was also introduced by J. Kelly et al. [5]. It's a point to point model that keeps a memory of the previous entered point. The model implements LSTM layer to overcome the vanishing gradient problem where gradient information disappears over time.
- Sequence to Point:
  Sequence to point was introduced by C. Zhang et al [6]. It's a CNN model where it maps a sequence of the input power to a midpoint in the sequence of the appliance power consumption.
- Recurrent network with GRU: The idea of using gate recurrent unit (GRU) instead of LSTM in recurrent network was proposed by D. Murray et al. [4], Krystalakos et al. [9] and [19]. Since the there is different models with different implementations, we will use the LSTM model from (b) but we replace the use of LSTM with GRU to medicate the poor performance of LSTM achieved by J.Kelly experiment.

A. Data Set Selection

The data set that was selected for this experiment are REDD [15] and UK-DALE [10] data set. These two data set was the most used data set from our literature review. Due to the different sampling of the two data set, we re-sampled the data set to 1 sample per 6 seconds. For UK-DALE, we used data from house #1 and house #2 to train our models while we used data from house #5 for testing. For REDD, we used data from house #1 and house #2 to train and data from house #3 for testing. Since UK-DALE data set has data of a period of more than 4 years and REDD data set has a period of around 3 months, we only selected a small portion window frame of UK-DALE that is roughly of around 6 months and its the same time window that was used by J. Kelly et al. [5] in their experiment. The Data sets were converted to a NILMTK [18] compatible format NILMTK [2] (non-intrusive load monitoring toolkit) is a python library that simplify extracting, processing and handling data from NILM data set.

B. Appliance Selection

We want to test our model of the two type of appliances, on/off state and the multi-state appliances. Also, the selected appliance needed to exist on both the data set to be used for the transferability test. Two appliance were selected for the experiment, microwave which represents on/off state appliance and the dish washer which represents the multi-state appliance.

[1] github.com/hyl0327/neuralnilmtp
[2] github.com/nilmtk/nilmtk

We wanted to include the washing machine to our test, but due to some issues we faced during code implementation that prevented us from extracting and performing data augmentation on washing machine data from REDD data set we had to exclude it from our test.

C. Data Augmentation

Here we prepared the data according to the experiment design of J. Kelly et al. [5]. Instead of taking a portion of main data and the a portion of matching time-frame from the appliance, we follow a complex procedure where we select the data by the activation of the appliance. We select all the activation of the desired appliance in our data set that satisfy the criteria in Table I. This insures that only complete activation event of an appliance is used for the experiment. These activations are then matched with the main power data that aligns with it. Finally, a random portions of the main power data are selected with the condition that the target appliance is not active during the selected time-frame.
Synthetic data were also used in the experiment. We created synthetic data by combining the activations of multiple appliances with the target appliance to create a new input data. This procedure was suggested by J. Kelly et al. [5] paper which helped increasing the amount of the data to be used for training that is according to our activation selection criteria. Synthetic data were only used for training and it was created by combining the activation of the following appliances: kettle, washing machine, dish washer, microwave and fridge. The code for performing data augmentation and synthesising was taken from J. Kelly et al. [5] github repository[3], although we had to modify it since it does not work anymore due to the incompatibility of the python version using in the J.Kelly project with the minimum version of the dependencies it needs.

TABLE I. CRITERIA FOR SELECTING ACTIVATION, THE SAME CRITERIA USED IN J.KELLY ET AL. [5] EXPERIMENT

Appliance	Max power (watts)	On power threshold (watts)	Min. on duration (secs)	Min. off duration (Sec)
Microwave	3000	200	12	30
Dish washer	2500	10	1800	1800

D. Data Normalization

There is different approaches for normalizing energy data. In C.Zhang et al. [6] experiment, they subtracted the input and the target data by the mean and then divides them by the standard deviation. While in J. Kelly et al. [5] experiment, they only divided the input by the standard deviation of a random subset from the whole input, while for the target data they divided it by it's maximum power draw. However, they updated their project code to divides the target data by a standard deviation of a randomly selected subset from the whole target data. A common approach when normalizing data for deep learning is to just divide both the input and the target values by the maximum value in the input data. We performed a quick test using these approaches on small portion of data, and we found that by using the standard deviation of a randomly selected subset of the input and the target and then dividing them by their the computed standard deviation achieved better results. Hence we selected this approach for our experiment.

[3]github.com/JackKelly/neuralnilm

E. Training

The models were implemented using TensorFlow with Keras and were trained on our personal machine with NVIDIA GTX 970. The window length of data used are depending on the appliance and is taken from J. Kelly et al. [5] experiment. For microwave we used a window length of 288 sample (1728 seconds) while the dish washer we reduced the window length to 1024 sample (6144 seconds) from 1536 sample (9126 seconds) due to getting out of memory error in our computer. Furthermore, the window length used for training the sequence to point model were reduced to 288 samples for both appliance, this was done to overcome the out of memory error since when training the sequence to point model we need to apply a sliding window approach on the input data and map each subset to a single point. The training epochs varies from model to model. For DAE we used epoch of 30, while for RNN LSTM and GRU we reduced the number of epoch to 20 due to it taking longer to train in our hardware. For the sequence to point model, we used an epoch of only 10, this was done because it takes roughly around 40 minute per a single epoch due to the increase in the amount of data when sliding over the input.

F. Evaluation Criteria

Each approach is trained on an on/off state appliance and on a multi-state appliance to showcase it's capabilities in not only detecting the simple patterns of on/off appliances, but to also detect the complex patterns of the multi-state appliances like the dish washer. Furthermore, the transferability test will showcase models capabilities in generalizing the learned features to other unseen, new instances of the same appliance. Each model will try to estimate the active power (measured in watts) of each appliance when given the active power of the total load. In this experiment, will be using the mean absolute error (MAE) to measure the estimation accuracy of each models under the different mentioned circumstances. The formula for MAE can be defined as follows:

$$MAE = \frac{\sum_{i=1}^{n} abs(y_i - x_i)}{n}$$

The sum of the of the differences between the prediction value (y_i) and the true value (x_i) is then divided by the number of samples (n).

G. Validity Evaluation

In our experiment, we want to conduct a transferability test on each trained model. Due to the differences in power consumption of each house in the two data set used, this will result in different computed standard deviation that is used for normalizing the data. To overcome this issue, the input data on both data sets were normalized using the same standard deviation. The standard deviation that is used are the average of the standard deviation of two randomly selected subsets of each data set. For the target appliance data, it was normalized in regular manner. although this will results in the appliance being normalized differently in each data set, we believe that this won't hurt the results of the appliance as the normalization will not affect it's power distribution (i.e.

assuming that microwave on REDD data set consumes 1100 watt and microwave on UK-DALE consume 1500 watts, when both are normalized that will be in the range of small numbers that are almost identical to each other. After prediction when we multiple the predicted number by the standard deviation that was used for normalizing it, it should results on both 1100 and 1500 values being brought back). Our experiment results unfortunately might not be generalized, since we only used a small portion of the data sets. This also accompanied by using low and uneven epoch per model during training. This was necessary due to the limited hardware power on the conducted computer which takes a long time to train these networks and the time constraint on our experiment. Another validity concern regarding the transferability test is that all our models does not have dropout layers, which helps in reducing overfitting. The decision to not add dropout layer was that the original architecture of each model from their respective research paper did not include dropout layers. Also since we are performing low number of epochs during training, we felt that using dropout layers could have negative effects to our results.

IV. Experimental Results and Analysis

In Table II we showcase the mean absolute error of microwave and dishwasher when trained on each models in REDD data set. While Table III shows the results on UK-DALE data set. The sequence to point model achieved the highest accuracy on both data set and both appliances. This can be attributed to the deep architecture that consists of 5 Convolution layers, which adds more parameter that can capture more feature on power patterns. Although it took the sequence to point the longest to train, we only trained it for 10 epochs compared to 30 of DAE and 20 to the RNN's models and we used a shorter window length compared to DAE and to the suggested window length by C.Zhang et al. [6] to coup with our shortage in GPU memory. The DAE model outperformed both recurrent notworks on both data sets and appliances, which confirms J. Kelly et al. [5] finding that CNN models outperforms RNN model. On the other hand, the use GRU instead of LSTM did improve the performance on on/off state machine (microwave) while significantly improving the performance on the multi-state appliance (dishwasher). Looking at the results from both data sets we can notice that UK-DALE produce a greater challenge than REDD data set to our models, as UK-DALE contains more appliances per house that makes the input power signal more noisy. In Fig. 1 we visualize some example disaggregations that was performed by the four models on both data sets.

TABLE II. The appliance mean absolute error (MAE) in watts for REDD data set. Best results are shown in bold.

Appliance	DAE	RNN LSTM	Seq2Point	RNN GRU
Microwave	26.39	42.04	**13.15**	34.58
Dish washer	51.02	90.76	**9.93**	62.77

Regarding the transferability test, the results can be seen in Table IV for training on REDD and testing on UK-DALE, and Table V and the average MAE of both when trained and tested on same data set or different can be seen in Fig. 2. We can see that on UK-DALE to REDD data sets, the sequence to point model achieves better results than the others. However, on the REDD to UK-DALE data sets it fails behind the DAE and RNN LSTM. Although sequence to point model achieved better results in same data set tested and when trained on UK-DALE and tested on REDD, it sufferers the biggest increase in MAE when trained on REDD and tested on UK-DALE. We believe it was due to REDD data set containing less data (in terms of time-frame window and number of activations for trained appliance) compared to UK-DALE, this and with the low number of epoch used for training the sequence to point model might had an effect in it's performance when trained and tested on different data sets. The low number of data on the REDD data set also affected other models when trained on REDD and tested on UK-DALE (compared the other way around). The RNN LSTM model seems to be the least affected by testing on different data set than the one trained on, followed up by RNN GRU. This could be an indication on the capability's of RNN in transferring well to other data sets. Overall, the sequence to point model still has the best average MAE on both tests despite having a massive increase in error when transferred to other data sets.

TABLE III. The appliance mean absolute error (MAE) in watts for UK-DALE data set. Best results are shown in bold.

Appliance	DAE	RNN LSTM	Seq2Point	RNN GRU
Microwave	39.61	57.12	**20.21**	46.64
Dish washer	61.17	93.18	**16.61**	65.73

TABLE IV. The appliance mean absolute error (MAE) in watts when trained on REDD data set and tested on UK-DALE data set. Best results are shown in bold.

Appliance	DAE	RNN LSTM	Seq2Point	RNN GRU
Microwave	**46.19**	56.96	66.80	59.98
Dish washer	152.69	**92.94**	100.78	148.35

TABLE V. The appliance mean absolute error (MAE) in watts when trained on UK-DALE data set and tested on REDD data set. Best results are shown in bold.

Appliance	DAE	RNN LSTM	Seq2Point	RNN GRU
Microwave	49.22	42.14	**41.97**	54.39
Dish washer	87.35	138.90	**78.60**	88.26

V. Conclusion

A brief background on energy disaggregation techniques were discussed. Then a brief summary of deep learning-based approaches using low frequency data was presented. We implemented four models selected from our research literature that we mentioned in our literature review. We applied the four models on two data sets, REDD and UK-DALE which are the two most used data sets in energy disaggregation research papers. We conducted an experiment were we trained the mentioned four models twice for each of the two data sets. The models were then tested on both the same data set (testing on unseen data from another house) and on the other data set. The comparison was carried out between the four models using mean absolute error on two scenarios, training and tested on the same data set and training and testing on two different data

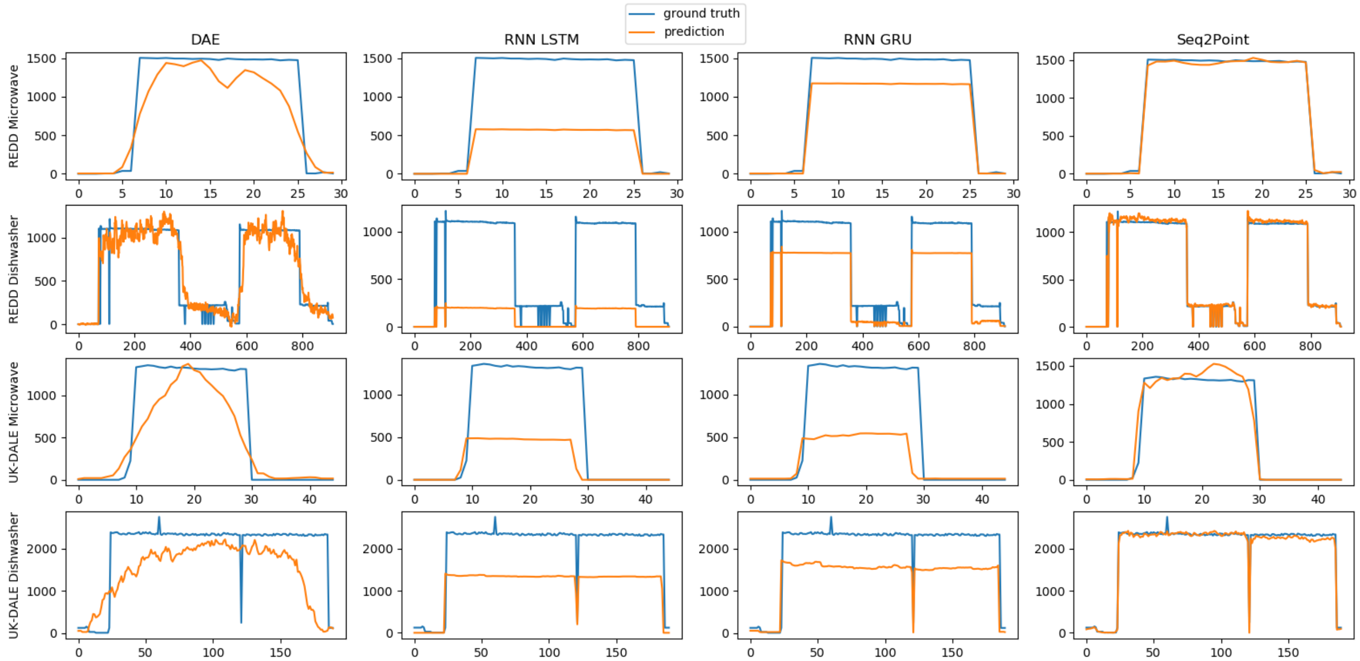


Fig. 1. Example of disaggregation results for microwave and dishwasher on both REDD and UK-DALE data set. The Y-axis corresponds to watts.

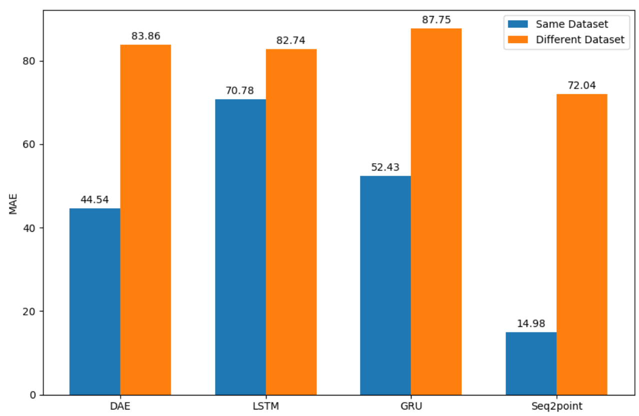


Fig. 2. The average MAE of each model when trained and tested on the same data set or on different data set.

sets. We noticed that the sequence to point model proposed by C.Zhang et al. [6] achieved the best results with an average MAE of 14.98 watts when tested on the same data set and 72.0 watts when transferred to a different data set. While on the other hand, both recurrent models performed the worst in the same data set testing and achieving close to the average score in the transferability test. Hence, the transferability is the most challenging issuing that is limiting the scalability of NILM-based solutions.

References

[1] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, Dec 1992.

[2] C. Fischer, "Feedback on household electricity consumption: a tool for saving energy?" *Energy Efficiency*, vol. 1, no. 1, p. 79–104, 2008.

[3] H. Rashid, P. Singh, V. Stankovic, and L. Stankovic, "Can non-intrusive load monitoring be used for identifying an appliance's anomalous behaviour?" *Applied Energy*, vol. 238, no. C, pp. 796–805, 2019. [Online]. Available: https://ideas.repec.org/a/eee/appene/v238y2019icp796-805.html

[4] D. Murray, L. Stankovic, V. Stankovic, S. Lulic, and S. Sladojevic, "Transferability of neural network approaches for low-rate energy disaggregation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 8330–8334.

[5] J. Kelly and W. Knottenbelt, "Neural nilm: Deep neural networks applied to energy disaggregation," in *Proceedings of the 2Nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, ser. BuildSys '15. New York, NY, USA: ACM, 2015, pp. 55–64. [Online]. Available: http://doi.acm.org/10.1145/2821650.2821672

[6] C. Zhang, M. Zhong, Z. Wang, N. H. Goddard, and C. A. Sutton, "Sequence-to-point learning with neural networks for nonintrusive load monitoring," in *AAAI*, 2016.

[7] M. Valenti, R. Bonfigli, E. Principi, and a. S. Squartini, "Exploiting the reactive power in deep neural models for non-intrusive load monitoring," in *2018 International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–8.

[8] G. Bejarano, D. Defazio, and A. Ramesh, "Deep latent generative models for energy disaggregation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 850–857, 2019.

[9] O. Krystalakos, C. Nalmpantis, and D. Vrakas, "Sliding window approach for online energy disaggregation using artificial neural networks," in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, ser. SETN '18. New York, NY, USA: ACM, 2018, pp. 7:1–7:6. [Online]. Available: http://doi.acm.org/10.1145/3200947.3201011

[10] J. Kelly and W. Knottenbelt, "The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes," *Scientific data*, vol. 2, p. 150007, 2015.

[11] L. Mauch and B. Yang, "A novel dnn-hmm-based approach for extracting single loads from aggregate power signals," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2384–2388.

[12] W. He and Y. Chai, "An empirical study on energy disaggregation via deep learning," in *2016 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE 2016)*. Atlantis Press, 2016/11. [Online]. Available: https://doi.org/10.2991/aiie-16.2016.77

[13] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

[14] S. Makonin, F. Popowich, L. Bartram, B. Gill, and I. V. Bajić, "Ampds: A public dataset for load disaggregation and eco-feedback research," in *2013 IEEE Electrical Power Energy Conference*, Aug 2013, pp. 1–6.

[15] J. Z. Kolter and M. J. Johnson, "Redd: A public data set for energy disaggregation research," in *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*, vol. 25, no. Citeseer, 2011, pp. 59–62.

[16] D. Murray, L. Stankovic, and V. Stankovic, "An electrical load measurements dataset of united kingdom households from a two-year longitudinal study," *Scientific Data*, vol. 4, no. 1, May 2017.

[17] C. Shin, S. Joo, J. Yim, H. Lee, T. Moon, and W. Rhee, "Subtask gated networks for non-intrusive load monitoring," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 1150–1157, 2019.

[18] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, "Nilmtk: An open source toolkit for non-intrusive load monitoring," 07 2014.

[19] P. Xiao and S. Cheng, "Neural network for nilm based on operational state change classification," *ArXiv*, vol. abs/1902.02675, 2019.

Pedestrian Crowd Detection and Segmentation using Multi-Source Feature Descriptors

Saleh Basalamah[1], Sultan Daud Khan[2]
Umm Al-Qura University, Saudi Arabia[1]
National University of Technology, Pakistan[2]

***Abstract*—Crowd analysis is receiving much attention from research community due to its widespread importance in public safety and security. In order to automatically understand crowd dynamics, it is imperative to detect and segment crowd from the background. Crowd detection and segmentation serve as pre-processing step in most crowd analysis applications, for example, crowd tracking, behavior understanding and anomaly detection. Intuitively, the crowd regions can be extracted using background modeling or using motion cues. However, these model accumulate many false positives when the crowd is static. In this paper, we propose a novel framework that automatically detects and segments crowd by integrating appearance features from multiple sources. We evaluate our proposed framework using challenging images with varying crowd densities, camera viewpoints and pedestrian appearances. From qualitative analysis, we observe that the proposed framework work perform well by precisely segmenting crowd in complex scenes.**

***Keywords*—Crowd detection; Fourier analysis; crowd analysis; crowd segmentation**

I. Introduction

With the growing population of the world and with the increased urbanization, scientific community focus on developing tools and techniques to ensure crowd safety. Events like sports, festivals, concerts, and carnivals, where the participants count in thousands, may lead to crowd disaster. Therefore, event organizers and security personnel must adopt adequate safety measures to ensure crowd safety. Crowd disasters still occur very frequently despite strict security and safety measures. One of the main reason of crowd disaster is the thousands of people gathered in a constrained environment which results in critically increased densities.

In order to ensure public safety in such high density situations, surveillance cameras are installed in multiple locations providing coverage of whole crowd scene. Generally, this is the job of security personnel to detect abnormal activities by watching over the TV. This kind of manual surveillance is a tiresome job and due to limited human capacity prone to human errors. Therefore, as solution an automatic analysis of the crowd is required which can reliably analyze the crowd dynamics. Designing such virtual analyst has attracted the interest of computer science community. Despite the recent advancements in computer vision technology, research community still did achieve desired result for understanding crowd dynamics. This attributes to following reasons: (1) Most of existing methods are based on assumptions often violated in real world scenarios. (2) Due to limited data availability, it is hard to train network effectively. Therefore as a solution, crowd simulation models have been used to provide synthetic data for training and also for validation of the computer vision algorithms. In order to automatically analyze the crowd dynamics, several computer vision tool sets [1], [2], [3], [4], [5], [6] are proposed. Using live video stream, these tools compute important measurements that are of significant importance to crowd managers and security personnel. These measurements include but not limited, crowd counting, density estimation, anomaly detection, crowd tracking. Although these tool sets computes important measurements that are useful in understanding crowd dynamics yet these tools could not detect detect crowds in the scene.

For understanding crowd dynamics, detection and tracking of pedestrians are the important steps. However, before starting any crowd analysis, crowd detection and segmentation is the preprocessing step. Intuitively , Crowd detection and segmentation can be achieve by motion segmentation techniques. But we observe that in real videos, large portion of crowd remain stationary and these stationary groups can be captured using motion segmentation techniques. Another shortcoming of motion segmentation is that it accumulates many false positives by detection motion of objects belong to other categories.

Crowd detection and segmentation serve as pre-processing step in many applications of crowd surveillance. However, crowd detection and segmentation is challenging task due to the following reasons. (1) In high density crowds, pedestrians generally stand close to each other due to constrained and limited space and environment. This cause severe occlusions among pedestrians. (2) Severe clutter in the scene usually confuse detector to distinguish between background and crowd.

In order to address above challenges, we proposed a framework that integrate appearance features and train a linear Support Vector Machine (SVM) classifier. Our proposed framework takes input of arbitrary size and divide into multiple cells in a grid from. Then for each cell we compute three descriptors, i.e, Local Binary Pattern (LBP), Fourier Analysis and Gray Level Co-occurrence Matrix (GLCM). The corresponding appearance features are then concatenated in a lineary fashion and SVM classifier is trained to classify each cell into crowd or non-crowd patch. Later on, we employ 11 x 11 2D-gaussian kernel to smooth the final output.

Our contributions can be summarized as follows:

- Our method does not require detection and tracking instead rely on low level features that works well in all crowd scenes.
- Our approach reduce the computational cost by detecting crowd from a single image instead of using whole video sequence.

- Our approach do not use background subtraction and do not make use of motion information.
- Our approach do not rely on pedestrian detection and tracking yet utilized appearance feature, therefore, can be applicable in both low and high density crowds.
- Our approach ease the process of crowd analysis by detecting only Region of Interest (crowded area).
- We evaluate our method on different scenes. The experiments results shows that our proposed method can precisely localize the crowd.

The rest of paper is organized as follows: we discuss related work in Section II, proposed methodology is discussed in Section III, Section V discusses experiment results and Section VI concludes the paper.

II. Related Work

There is inadequate work reported in literature on crowd detection segmentation. Most of crowd related literature focus on crowd counting, density estimation, tracking and anomaly detection. Automated analysis of crowd behaviors has large number of applications applications ranging from prediction of congestion to the discovery of abnormal behaviors or flows. Most of the research is focused on detecting anomalies in videos [7], [8], [9], [10], [11], [12], [13], counting people in crowds [14], [15], [16], [17], [18], [19], [20], [21], characterizing different motion flows and segmentation [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32]. Other works aimed at detecting/tracking individuals or group of individuals in crowd scene [33], [34], [35], [36], [37], [38], [39], [40] rather than identifying crowd behaviours. As there is growing interest in crowd dynamics understanding in general, identifying crowd behaviors specifically has not been studied in depth, and very few papers have explicitly concentrated on identifying or modelling crowd behaviour. Berkan et al. [41] identify five crowd behaviours, i.e. blocking, lane, bottleneck, ring/arch, fountainhead. A similar framework is proposed by [42], where the same five crowd behavior are identified. Khan et al. [43] proposed a method that can identify crowd behaviors by utilizing source and sink information of trajectories. Recognizing the recent success of convolution neural network in filed of object detection and segmentation, [44] proposed CNN based on two stream network fusion network [45], originally designed for video action recognition to identify multiple crowd behaviors in crowd scene. However, Convolution neural network (CNN) has not gained adequate performance that have been achieved in image classification and object detection. Part of reason is the lack of data set for training or the data sets are too small and noisy. Compared to the image classification, classification of crowd behaviors has the additional challenge of variations in motion, viewpoint and scales. Due to these challenges, we require more training examples. Another reason is that CNN are not able to take full advantage of temporal information existing between consecutive frames of the video.

III. Proposed Methodology

The overall methodology of our proposed framework is shown in Fig. 1. During training phase, given a set of images, we first divide each image into grid of cells. Then for each cell we compute low-level appearance features. including local-binary pattern, Fourier analysis and Gray Level Co-occurrence Matrix (GLCM). We then construct a long feature vector by concatenating all features. Then a linear classifier is trained using long feature vector. The size of feature vector is 128. During testing phase, each input image passed through the same steps and extracted long feature vector is then mapped to learned classifier for generating the confidence score for each cell. Later on, Gaussian kernel is applied to smooth the final output.

A. Local Binary Pattern

For texture and appearance base classification, Local Binary Pattern (LBP) is a perfect choice. For each pixel p, we compute LBP with angular quantization of 8 pixels and with spatial resolution of 1. We then compare pixel p with its 8 neighbourhood pixels in such a way that the output is 0 if the intensity of pixel p is less than its neighbourhood and 1 if it is greater. We then process all pixels of input image in the same way and generate normalized histogram. With this unique representation, we can only capture the appearance information of the whole image. For capturing the texture information, we compute Gray-Level Co-occurrence matrix from the intensity values. We then compute Entropy (**P**), Energy (**E**), Contrast (**C**) and Homogeneity (**H**) as in the following equations.

$$P = -\sum_x \sum_y p[x,y] \log p[x,y] \tag{1}$$

$$E = -\sum_x \sum_y p^2[x,y] \tag{2}$$

$$C = -\sum_x \sum_y (x-y)^2 p[x,y] \tag{3}$$

$$H = -\sum_x \sum_y p[x,y]/1 + |x-y| \tag{4}$$

where x and y represent horizontal and vertical components of the input image.

B. Fourier Analysis

Fourier Analysis is a unique way of extracting appearance information from the image. We observed that in high density crowds, where large number of people gather in a constrained environment, far away pedestrians cover only few pixels due to perspective distortions. In this way, pedestrian detection and histogram of gradient can can impart useful information. We further observed that high density crowds have repetitive structures, as all pedestrians appear same from a distant camera view point. This unique appearance of crowd can be easily represented by Fourier Transform F_t. When we convert input image to Fourier domain, these repetitive structure of pedestrian heads can be easily detected using the peaks of frequency domain. In order to tackle scale problem that is caused by perceptive distortions, we divide the image into patches. Here, we assume that the density in the patch is same.

For each patch P that belongs to input image I, we perform the following steps:

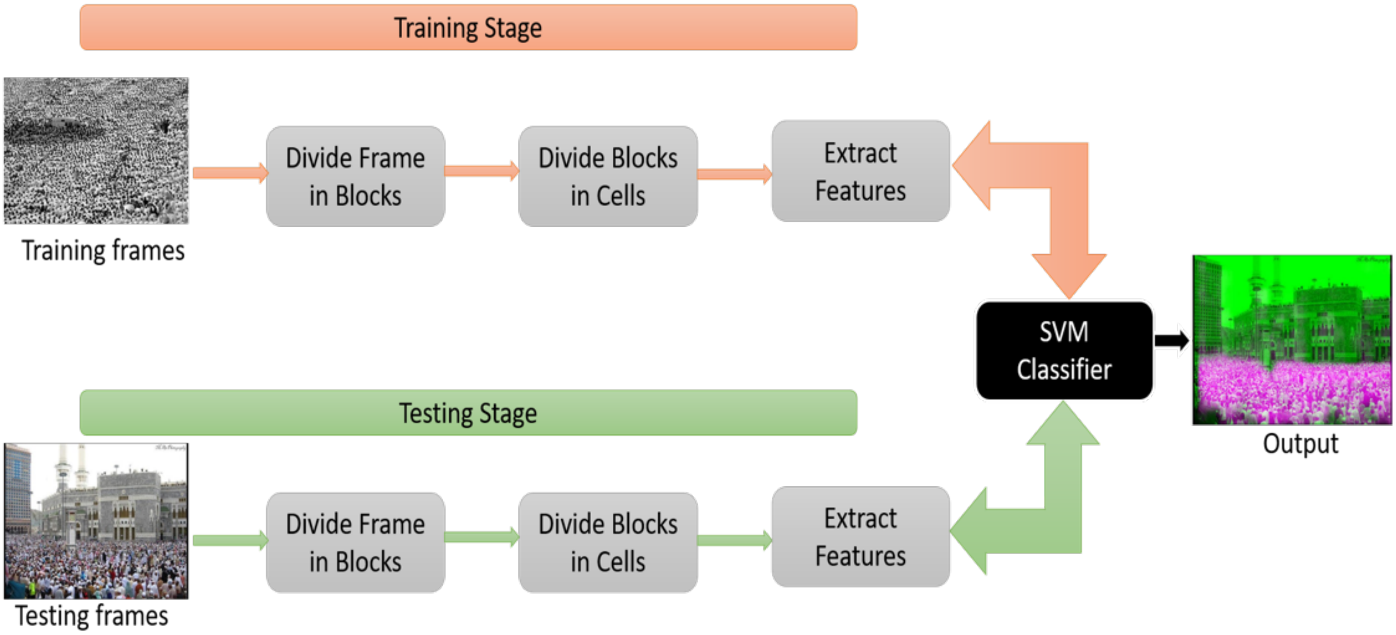


Fig. 1. Pipeline of proposed framework both during testing and training phase. Input image is divided into block and the into cells. Features are then extracted from each cell and train SVM classifier.

- Convert the patch into gradient Δ_P.
- Apply Fourier Transform on Δ_P and then apply low pass filter. This step is important to remove high frequency components from the signal as it contains information about the edges.
- Remove low amplitude component by applying a threshold υ. We set the value of υ to 0.4.
- Reconstruct the image P_r by applying inverse Fourier Transform and apply non-maxima suppression.

After reconstruction, we compute the following statistical features, mean (**M**), Variance (**V**), Skewness (**S**), and Kurtosis (**K**)

$$M = \frac{1}{xy} \sum_{x,y \in P_r} P_r(x,y) \tag{5}$$

$$V = \frac{1}{xy-1} \sum_{x,y \in P_r} \left(P_r(x,y) - \frac{1}{xy-1} \sum_{x,y \in P_r} P_r(x,y) \right)^2 \tag{6}$$

$$S = \frac{\frac{1}{xy-1} \sum_{x,y \in P_r} \left(P_r(x,y) - \frac{1}{xy-1} \sum_{x,y \in P_r} P_r(x,y) \right)^3}{\left(\frac{1}{xy-1} \sum_{x,y \in P_r} \left(P_r(x,y) - \frac{1}{xy-1} \sum_{x,y \in P_r} P_r(x,y) \right)^2 \right)^{\frac{3}{2}}} \tag{7}$$

$$K = \frac{\frac{1}{xy-1} \sum_{x,y \in P_r} \left(P_r(x,y) - \frac{1}{xy-1} \sum_{x,y \in P_r} P_r(x,y) \right)^4}{\left(\frac{1}{xy-1} \sum_{x,y \in P_r} \left(P_r(x,y) - \frac{1}{xy-1} \sum_{x,y \in P_r} P_r(x,y) \right)^2 \right)^2} \tag{8}$$

where P_r is image patch and x and y are the horizontal and vertical coordinates of the patch.

C. Gray Level Co-occurrence Matrix

Gray Level Co-occurrence Matrix commonly used for texture extraction and detection. GLCM uses distribution of gray-level of neighboring pixels in relation with center pixel. Marana et al. proposed a method of GLCM for utilizing texture information for crowd density estimation. In our case, we adapt it to train a binary classifier that distinguish background from the crowded patches. The Gray-Level Co-Occurrence Matrix (GLCM) $P[x,y]$ is computed by taking the sum of all pixel pairs having gray value x and y separated by distance parameter d and at an orientation $theta$ of 0sidegree 45sidegree, 90sidegree and 135sidegree, respectively. The counting is converted to joint conditional probability $P(x,y/\theta,d)$.

After computing GLCM for input image, we then extract features, for example, Entropy (**P**), Energy (**E**), Contrast (**C**) and Homogeneity (**H**) using equations 1, 2, 3, and 4.

IV. Features Fusion

In this section, we discuss the fusion of different features from different sources as shown in Fig. 2. For training, we use N number of frame. In order to ensure patch wise training, we divide each image into dense overlapping patches. We then extract appearance features using above mentioned sources and

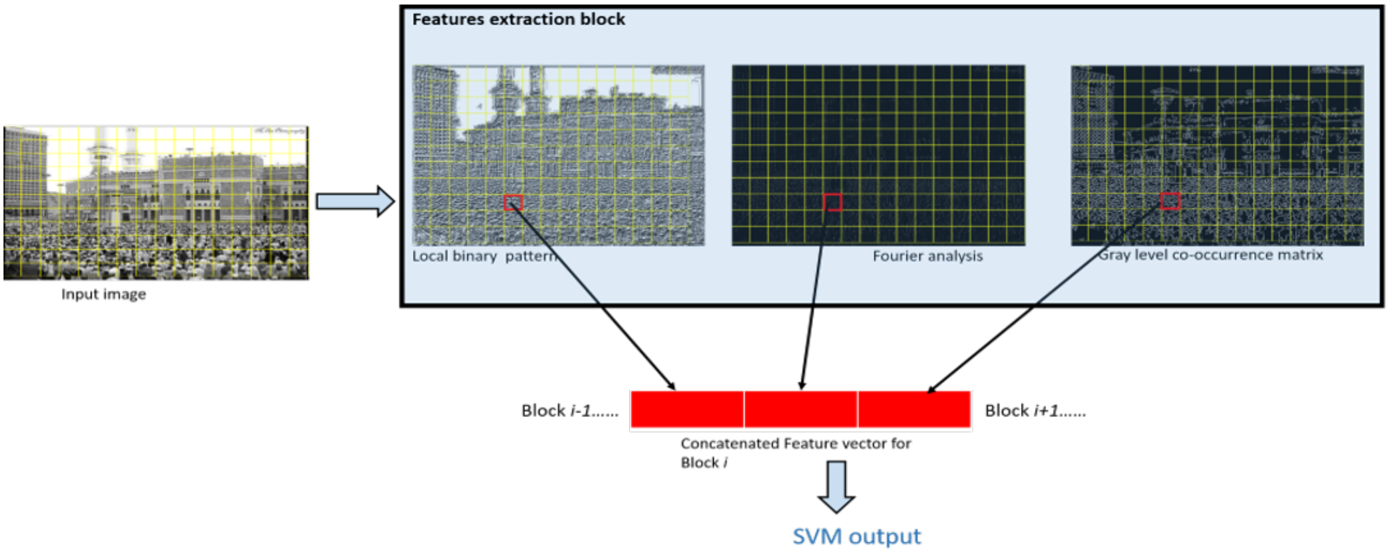


Fig. 2. Sample input frame is divided into multiple blocks. Feature are extracted for block *i*(red). Features are concatenated and feed to the SVM classifier to assign a score.

combine the resultant features into a long feature vector. Let $F_i = \{m_i^1, m_i^2 \ldots, x_i^M\}$ is the final feature vector for input image i and M represent the number of patches in the image.

V. Experimental Results

In this section, we evaluate the effectiveness of our proposed approach. In order to evaluate our proposed method, we publicly available UCF_CC_50 [16] dataset. UCF_CC_50 is challenging dataset that contains 50 images captured from 50 different scenes with significant variations in resolution, camera view points and densities. The density in images varies from 94 person / image to 4543 persons/image. We randomly divide the data set into training and testing samples using the same convention used in [16]. During training, we cropped multiple patches from the image and divide them into crowd and non-crowd patches. Fig. 3 clearly illustrates the input image with corresponding crowd and non-crowd patches. We feed these patches to train our classifier.

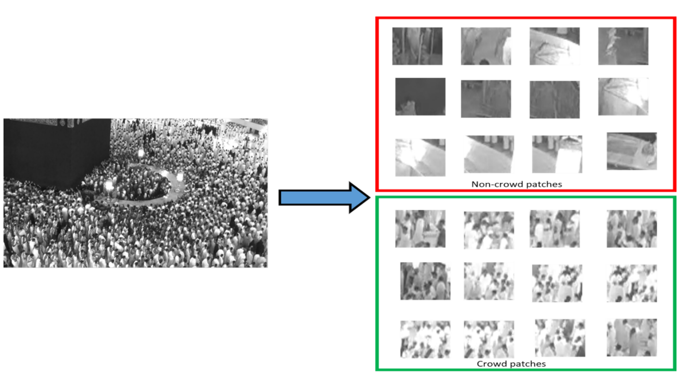


Fig. 3. Sample image used for training. The image is divided into positive and negative patches. We extract features from these patches and feed them for training SVM classifier.

The results of our proposed framework is shown in Fig. 5.

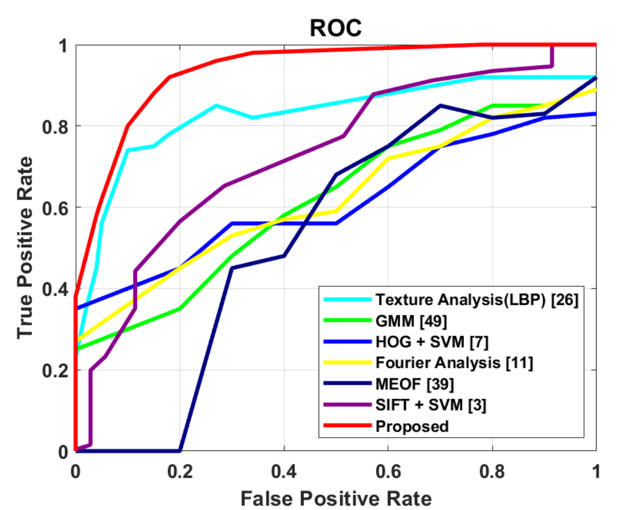


Fig. 4. ROC curves of different methods using UCF_CC_50 [16] dataset.

From the Figure, it is obvious that our proposed framework effectively discriminate crowd from non-crowd regions and precisely segment the crowded area. From the Fig. 5, it is obvious that our proposed method achieve impressive results with small number of false positives. These false positives attribute to the fact that our proposed framework also treat "leafy" areas as crowded regions.

We also compare our results with other baseline methods in a quantitative way. The first based line method is traditional Gaussian mixture model (GMM) for background subtraction. The second baseline model is based on motion extraction using optical flow (MEOF). Third method (HOG + SVM) is patch based model that utilizes Histogram of Oriented Gradient (HOG) features to train SVM classifier. In addition we also compare our results with SIFT + SVM [46], Fourier Analysis [16], and Texture [47]. We keep the size of patch to

TABLE I. COMPARATIVE ANALYSIS WITH OTHER TECHNIQUES ON UCF_CC_50 [16]

Method	**AUC**
GMM [48]	0.27
MEOF [27]	0.15
HOG + SVM [49]	0.45
SIFT + SVM [46]	0.56
Fourier analysis [16]	0.37
Texture analysis (LBP) [47]	0.58
Proposed	0.63

32 x 32 pixels in all our experiments. We use Area-under-curve (AUC) from ROC curves as evaluation metrics. We report the ROC and AUC results in Fig. 4 and Table I respectively. From the results, it is obvious that our proposed framework outperforms other state-of-the-art methods. We further observed that texture and appearance features work well in high density crowds since they capture regular and repetitive structure of the crowds.

VI. CONCLUSION

In this paper, we propose a novel approach to detect and segment crowd based on low-level appearance features. The proposed framework is tested on challenging images with large scale variations in densities, viewpoints and appearances of pedestrians in crowd. From experiment results, we showed that our proposed framework can precisely detect and segment crowded regions in the scene.

In future, we plan to integrate the proposed framework with various crowd tracking and crowd behavior understanding applications.

ACKNOWLEDGMENT

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU for this research.

REFERENCES

[1] S. D. Khan, M. Tayyab, M. K. Amin, A. Nour, A. Basalamah, S. Basalamah, and S. A. Khan, "Towards a crowd analytic framework for crowd management in majid-al-haram," *arXiv preprint arXiv:1709.05952*, 2017.

[2] S. D. Khan, M. Saqib, and M. Blumenstein, "Towards a dedicated computer vision tool set for crowd simulation models," *arXiv preprint arXiv:1709.02243*, 2017.

[3] S. D. Khan, G. Vizzari, and S. Bandini, "A computer vision tool set for innovative elder pedestrians aware crowd management support systems." in *AI* AAL@ AI* IA*, 2016, pp. 75–91.

[4] ——, "Facing needs and requirements of crowd modelling: Towards a dedicated computer vision toolset," in *Traffic and Granular Flow'15*. Springer, 2016, pp. 377–384.

[5] S. D. Khan, L. Crociani, and G. Vizzari, "Integrated analysis and synthesis of pedestrian dynamics: First results in a real world case study," *From Objects to Agents*, 2013.

[6] ——, "Pedestrian and crowd studies: Towards the integration of automated analysis and synthesis."

[7] H. Ullah, A. B. Altamimi, M. Uzair, and M. Ullah, "Anomalous entities detection and localization in pedestrian flows," *Neurocomputing*, vol. 290, pp. 74–86, 2018.

[8] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," 2009.

[9] H. Ullah and N. Conci, "Crowd motion segmentation and anomaly detection via multi-label optimization," in *ICPR workshop on Pattern Recognition and Crowd Analysis*, 2012.

[10] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1975–1981.

[11] H. Ullah, M. Ullah, and N. Conci, "Real-time anomaly detection in dense crowded scenes," in *Video Surveillance and Transportation Imaging Applications 2014*, vol. 9026. International Society for Optics and Photonics, 2014, p. 902608.

[12] H. Ullah, L. Tenuti, and N. Conci, "Gaussian mixtures for anomaly detection in crowded scenes," in *Video Surveillance and Transportation Imaging Applications*, vol. 8663. International Society for Optics and Photonics, 2013, p. 866303.

[13] H. Ullah, M. Ullah, H. Afridi, N. Conci, and F. G. De Natale, "Traffic accident detection through a hydrodynamic lens," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2470–2474.

[14] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 705–711.

[15] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 833–841.

[16] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2547–2554.

[17] M. Arif, S. Daud, and S. Basalamah, "Counting of people in the extremely dense crowd using genetic algorithm and blobs counting," *IAES International Journal of Artificial Intelligence*, vol. 2, no. 2, p. 51, 2013.

[18] ——, "People counting in extremely dense crowd using blob size optimization," *Life Science Journal*, vol. 9, no. 3, pp. 1663–1673, 2012.

[19] M. Saqib, S. D. Khan, and M. Blumenstein, "Texture-based feature mining for crowd density estimation: A study," in *Image and Vision Computing New Zealand (IVCNZ), 2016 International Conference on*. IEEE, 2016, pp. 1–6.

[20] S. Khan, G. Vizzari, S. Bandini, and S. Basalamah, "Detecting dominant motion flows and people counting in high density crowds," 2014.

[21] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," in *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*. IEEE, 2017, pp. 1–7.

[22] S. D. Khan, G. Vizzari, and S. Bandini, "Identifying sources and sinks and detecting dominant motion patterns in crowds," *Transportation Research Procedia*, vol. 2, pp. 195–200, 2014.

[23] M. Saqib, S. D. Khan, N. Sharma, and M. Blumenstein, "Extracting descriptive motion information from crowd scenes," in *2017 International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, 2017, pp. 1–6.

[24] M. Saqib, S. D. Khan, and M. Blumenstein, "Detecting dominant motion patterns in crowds of pedestrians," in *Eighth International Conference on Graphic and Image Processing (ICGIP 2016)*, vol. 10225. International Society for Optics and Photonics, 2017, p. 102251L.

[25] H. Ullah, M. Ullah, and N. Conci, "Dominant motion analysis in regular and irregular crowd scenes," in *International Workshop on Human Behavior Understanding*. Springer, 2014, pp. 62–72.

[26] S. Ali and M. Shah, "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–6.

[27] H. Ullah and N. Conci, "Structured learning for crowd motion segmentation," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 824–828.

[28] H. Ullah, M. Ullah, and M. Uzair, "A hybrid social influence model for pedestrian motion segmentation," *Neural Computing and Applications*, pp. 1–17, 2018.

[29] M. Gao, J. Jiang, J. Shen, G. Zou, and G. Fu, "Crowd motion segmentation and behavior recognition fusing streak flow and collectiveness," *Optical Engineering*, vol. 57, no. 4, p. 043109, 2018.

[30] H. Ullah, M. Uzair, M. Ullah, A. Khan, A. Ahmad, and W. Khan, "Density independent hydrodynamics model for crowd coherency detection," *Neurocomputing*, vol. 242, pp. 28–39, 2017.

[31] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu, "Unified crowd segmentation," in *European conference on computer vision*. Springer, 2008, pp. 691–704.

[32] H. Mansour, C. Dicle, D. Tian, M. Benosman, and A. Vetro, "Method and system for segmenting pedestrian flows in videos," Feb. 13 2018, uS Patent 9,892,520.

[33] G. J. Brostow and R. Cipolla, "Unsupervised bayesian detection of independent motion in crowds," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 594–601.

[34] L. Kratz and K. Nishino, "Tracking with local spatio-temporal motion patterns in extremely crowded scenes," 2010.

[35] S. D. Khan, G. Vizzari, S. Bandini, and S. Basalamah, "Detection of social groups in pedestrian crowds using computer vision," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2015, pp. 249–260.

[36] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *European conference on computer vision*. Springer, 2008, pp. 1–14.

[37] S. D. Khan, F. Porta, G. Vizzari, and S. Bandini, "Estimating speeds of pedestrians in real-world using computer vision," in *International Conference on Cellular Automata*. Springer, 2014, pp. 526–535.

[38] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 261–268.

[39] S. D. Khan, "Estimating speeds and directions of pedestrians in real-time videos: A solution to road-safety problem," in *CEUR Workshop Proceedings*, 2014, p. 1122.

[40] T. Zhao and R. Nevatia, "Tracking multiple humans in crowded environment," in *null*. IEEE, 2004, pp. 406–413.

[41] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 10, pp. 2064–2070, 2012.

[42] M. Ullah, H. Ullah, N. Conci, and F. G. De Natale, "Crowd behavior identification," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1195–1199.

[43] S. D. Khan, S. Bandini, S. Basalamah, and G. Vizzari, "Analyzing crowd behavior in naturalistic conditions: Identifying sources and sinks and characterizing main flows," *Neurocomputing*, vol. 177, pp. 543–563, 2016.

[44] C. Dupont, L. Tobías, and B. Luvison, "Crowd-11: A dataset for fine grained crowd behaviour analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 9–16.

[45] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

[46] O. Arandjelovic, "Crowd detection from still images," in *BMVC 2008: Proceedings of the British machine vision association conference 2008*. BMVA Press, 2008, pp. 1–10.

[47] R. Ma, L. Li, W. Huang, and Q. Tian, "On pixel count based crowd density estimation for visual surveillance," in *IEEE Conference on Cybernetics and Intelligent Systems, 2004.*, vol. 1. IEEE, 2004, pp. 170–173.

[48] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 2. IEEE, 2004, pp. 28–31.

[49] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005.

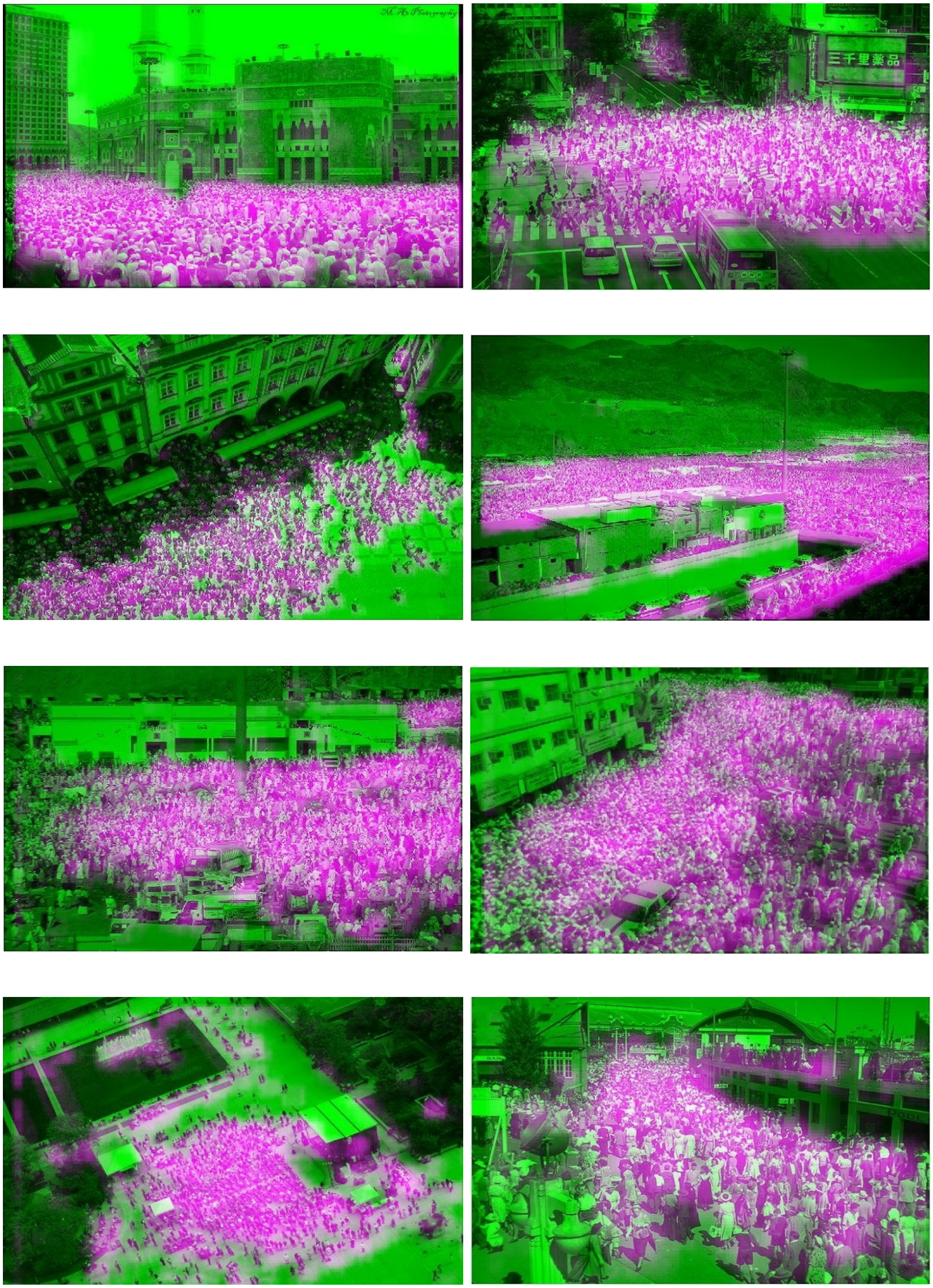


Fig. 5. Crowd segmentation results predicted by proposed approach in different crowd scenes: Segmentation mask is overlaid on image where the green color shows the background or non-crowd while pink color shows crowded areas).

Towards an Improvement of Fourier Transform

Khalid Aznag[1], Toufik Datsi[2], Ahmed El Oirrak[3], Essaid El Bachari[4]
Computer Systems Engineering Laboratory
Faculty of Sciences Semlalia
Marrakech, Morocco

***Abstract*—With the development of information technology and the coming period of large data, the image signals play an increasingly more significant role in our life because of the phenomenal development of system correspondence innovation, and the comparing high proficiency image handling strategies are requested earnestly. The Fourier transform is an important image processing tool, which is used in a wide range of applications, such as image filtering, image analysis, image compression and image reconstruction. It 's the simplest among the other transformation method used in mathematics. The real time consumption is lesser due to this method. It has a vast use in image processing, particularly object 2D, 3D and other representation. This paper proposes a new Fourier transform which is called Non Uniform Fourier Transform (NUFT). The proposed descriptor takes into consideration the change of point index. Also, an application is made on 2D set of points and a real image. The main advantages of the proposed transform are invariance under change of index point and robustness to noise. Also, the extraction of invariant under rotation and affinity is immediate because the linearity is assured. The proposed descriptor is tested on MPEG 7 database and compared with the normal Fourier transform to shows its efficiency. The experimental results prove the effectiveness of the proposed descriptor.**

Keywords—Fourier transform; NUFT; noise; invariant

I. Introduction

Fourier transform is an interesting image processing tool which used to decompose an image into sine and cosine components. Fourier transform is used in several fields such as image processing and filters, transformation, representation, etc. Historically, one of the most widely used shape description methods is Fourier descriptors (FD) [1], [2], [3], [4]. The discrete Fourier transform (DFT) is one of the most fundamental and important numerical algorithms which plays a central role in the image processing area, including image denoising [5], image feature extraction [6], and compressed sensing [7]. The Fast Fourier Transform (FFT) [8] which computes the DFT of an n-size signal in O(n log n) time greatly simplifies the complexity of DFT and gets a wide range of applications. In [9], [10] and [11], [12] are implementations of the discrete Fourier transform (DFT). Fourier transform hypotheses of central importance in a vast range of applications in engineering, applied mathematics, and physical science. In addition, Fourier transform is a mathematical concept which suited extremely well for signal analysis. In [13], [14] authors proposed sparse fast Fourier transform for one-dimension (1D-SFFT) signal which is faster than traditional DFT. However, the two-dimensional image signal is more broadly used, and a two-dimensional sparse Fourier transform cannot simply be constructed with a one-dimensional sparse Fourier transform. Sheng Shi et al in [15] proposed a new fast two-dimensional Fourier transform (2D-SFFT) that takes advantage of image sparsity.

In this work, the advantage of our method is that it is systematic and it allows obtaining the analytical form of all invariant polynomials of a given order, which was not the case using the Suck method and Flusser [16].

The rest of this paper will be organized as follows: Section II provides an overview of related work. Section III describes normal Fourier descriptors and the proposed Fourier transform. In Section IV the experiments are presented. The results and discussions are presented in Section V. Section VI contains the conclusion and the future work highlighted in Section VII.

II. Related Works

Fourier transform is one of the oldest and well-known methods in the field of the mathematics. It used in a wide range of applications, such as image filtering, image analysis, image reconstruction, and image compression. In the literature, several papers have described methods for approximating one-dimensional Non uniform Fast Fourier Transforms by interpolating an oversampled Fast Fourier transforms, start with [17] and including [18], [19], [20], [21], [22], [23], [24], [25], [26], [27]. In [28] the fundamentals of Fourier transform, Fourier series, discrete Fourier transform and fast Fourier transform with simple examples and review of Fourier transform to supply a clear understanding of its applications in power quality issues. In [29], the author presents a novel method for improving the Fast Fourier Transform (FFT) based spectral estimation for the diagnosis of faults in induction motors. In [30], the authors propose a new method for optical image encryption using fractional Fourier transform (FRFT). Two-dimensional fast Fourier transform (2D-FFT) is successfully applied to analyse images [31]. Image compression technique that uses real Fourier transform is proposed by Kekre et al in [32]. Their technique is applied on the image in three ways: Row transform, Column transform, and Full transform. Aznag et al. in [33] applied this new descriptor for 3D parameterized point set and 3D curve. Previous research has demonstrated that there is a non-uniform Fourier transform [16], [34], but the problem treated in this new work is not the spacing between samples, but simply the change in order of points in storage or manipulation of those points.

III. Methodology

This section discusses the proposed descriptor. We first give an overview of the normal Fourier descriptor. Then, we describe more in detail our proposed Fourier descriptor.

A. Normal Fourier Descriptor

1) Fourier descriptor for two-dimensional indexed point set: To define Fourier descriptor (FD) for 2D indexed point set; Let X (x(n), y(n)) n=1...N, denotes a closed contour with N is the number of points on the normalized contour and 2π as length, then the Fourier descriptors are given by equations. (1), (2) and (3).

$$Z(k) = \begin{pmatrix} u(k) \\ v(k) \end{pmatrix} \quad (1)$$

With

$$u(k) = \frac{1}{N}\sum_{n=1}^{N} x(n) e^{-j\frac{2\pi kn}{N}} \quad (2)$$

$$v(k) = \frac{1}{N}\sum_{n=1}^{N} y(n) e^{-j\frac{2\pi kn}{N}} \quad (3)$$

For $k = 1 \dots N$

Where u and v represent the Fourier descriptors of x and y respectively.

2) Shift theorem and transformation effect: Let X and $\tilde{X}$ are two objects having the same shape but with a shift in starting points, then (Equations. (4), (5) and (6)):

$$\tilde{Z}(k) = e^{jkl_0} AZ(k) + B\delta_k \quad (4)$$

For all integer k,

$$Z(k) = \big(u(k), v(k)\big) \quad (5)$$

and

$$\tilde{Z}(k) = \big(\tilde{u}(k), \tilde{v}(k)\big) \quad (6)$$

Are respectively the bi-dimensional Fourier descriptor vectors of X and $\tilde{X}$.

Where

$$\delta_k = \begin{cases} 1 & if \quad k = N \\ 0 & otherwise \end{cases} \quad (7)$$

Is the Kronecker symbol (Equation. (7)). The real l_0 denotes the difference between starting points on a contour and its transformed.

B. Proposed Fourier Descriptor

In this section, we present the proposed descriptor. As normal Fourier transform is dependent on point index. So, to solve this problem we define the novel descriptor. We apply the proposed Fourier transform for the 2D parameterized point set and the binary objects. The idea of our descriptor is simple and easy, we replace integer k by parameter τ (see Equation. (16)).

In general, we have two cases:

- A structured set of points, indexed by n integer which represents the order of points.
- Unstructured set of points, in this case the order is not respected and we propose the use of another characteristic of a point, which is independent of the order.

3) Novel Fourier descriptor for two-dimensional indexed point set: To define novel Fourier descriptor (FD) for 2D indexed point set; Let $X(\boldsymbol{\tau}) = (\mathbf{x}(\boldsymbol{\tau}), \mathbf{y}(\boldsymbol{\tau}))$ denotes a 2D set of point having $\boldsymbol{\tau}$ as parameter and N is the number of points, then the novel 2D Fourier descriptors is defined by.

$$\begin{cases} u(k) = \frac{1}{N}\sum_{\tau} x(\tau) e^{-j\frac{k\tau}{T}} \\ v(k) = \frac{1}{N}\sum_{\tau} y(\tau) e^{-j\frac{k\tau}{T}} \end{cases} \quad (8)$$

Where

$$T = \int \tau d\tau$$

4) Linearity: Let $X = (x(\tau), y(\tau))$ and $\tilde{X} = (\tilde{x}(\tau), \tilde{y}(\tau))$ denote two shapes having the same shape with an affine transform difference i.e (Equations. (9), (10), (11), (12), (13) and (14)):

$$\tilde{X} = AX \quad (9)$$

Where

$$A = \begin{pmatrix} a_{11} a_{12} \\ a_{21} a_{22} \end{pmatrix} \quad (10)$$

So

$$\tilde{x} = a_{11}x(\tau) + a_{12}y(\tau) \quad (11)$$

$$\tilde{y} = a_{21}x(\tau) + a_{22}y(\tau) \quad (12)$$

Then

$$\tilde{u}(\tau) = \frac{1}{N}\sum_{n=1}^{N} \tilde{x}(\tau) e^{-\frac{jk\tau}{T}} \quad (13)$$

$$= \frac{1}{N}\sum_{n=1}^{N} \big(a_{11}x(\tau) + a_{12}y(\tau)\big) e^{-\frac{jk\tau}{T}}$$

$$= a_{11}u(k) + a_{12}v(k)$$

The same for:

$$\tilde{v}(k) = a_{21}u(k) + a_{22}v(k) \quad (14)$$

In case of rotation transformation, a 2×2 transformation matrix is defined by

$$A = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix} \quad (15)$$

Where θ is the angle of rotation. This principle of linearity allows us to extract invariant easily from shapes X and $\tilde{X}$.

IV. Experiments

A. Application on Sets of Points

In these experiments, we have four sets of 2D points with x and y coordinates shown in Fig. 1. We present the experimental test of our proposed descriptor in four shapes that are shown in Fig. 1(a), 1(b), 1(c) and 1(d). Using Equation (16), the parameter τ used in this experiment is defined by

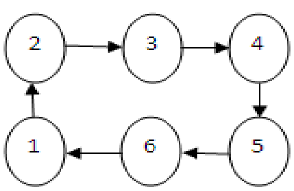
$$\tau = \sqrt{x^2 + y^2} \tag{16}$$

B. Application on Real Image

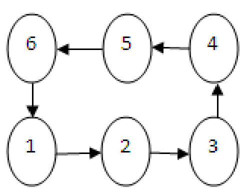
To test our approach; we apply our method on the MPEG7 database [35]. In this database there are 70 classes of shapes, each one has 20 members as shown in Fig. 6. In this section we applied our approach to a real image, an MPEG7 image of size 750×531 is presented in Fig. 2. To obtain contour points we can browse edge image using row by row or column by column. So, the coordinates of contours have not the same index using previous browses. Starting points on each contour are presented by a circle (as shown in Fig. 3). Not only the starting points are different but also the order of other points. Fig. 4 shows results with contours of an elephant.

C. Noise Effect

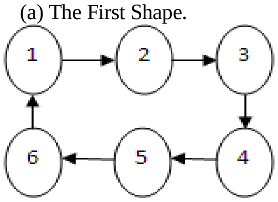
In order to test our new Fourier transform for noise, we add noise to the Elephant image (see Fig. 5). The percentage of added noise is 10% and we take the first 10 coefficients (development into Novel transformed of coordinates x).



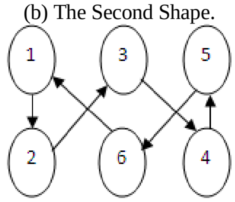
(a) The First Shape.



(b) The Second Shape.



(c) The Third Shape.



(d) The Fourth Shape.

Fig. 1. A Set of Points with Associated 2D Shapes.

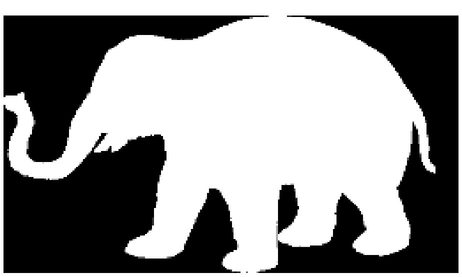
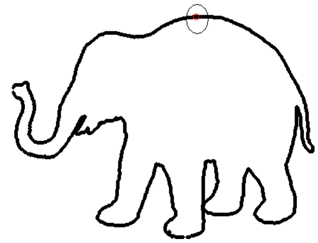
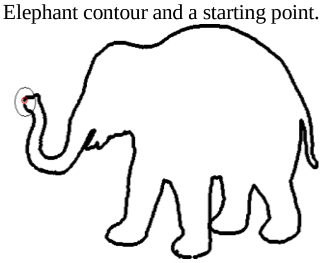


Fig. 2. Elephant Image.

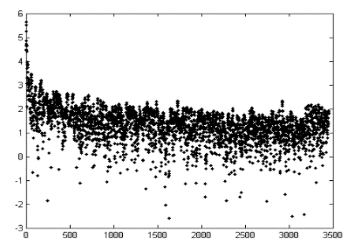


(a) Elephant contour and a starting point.

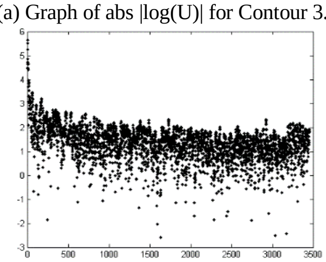


(b) Elephant contour with another starting point.

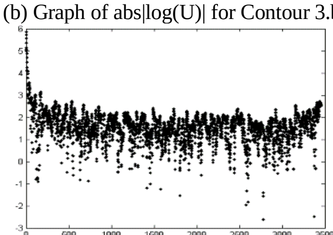
Fig. 3. The Contours of Elephant Represent the Same Content but the Order of Points is not the Same.



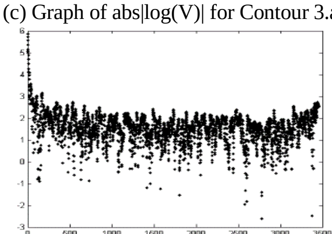
(a) Graph of abs |log(U)| for Contour 3.a)



(b) Graph of abs|log(U)| for Contour 3.b)

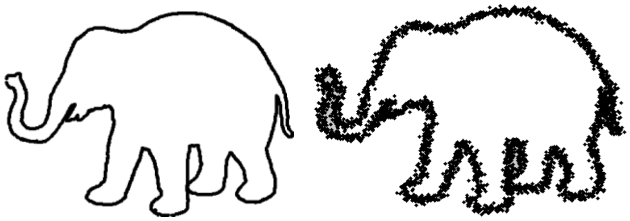


(c) Graph of abs|log(V)| for Contour 3.a)



(d) Graph of abs|log(V)| for Contour 3.b)

Fig. 4. Graphs Represent Abs |Log(U)| (a) and (b), Graphs Represent Abs |Log(V)| (c) and (d).



(a) Original elephant image. (b) Noisy elephant image.

Fig. 5. Original and Noisy Elephant Images.

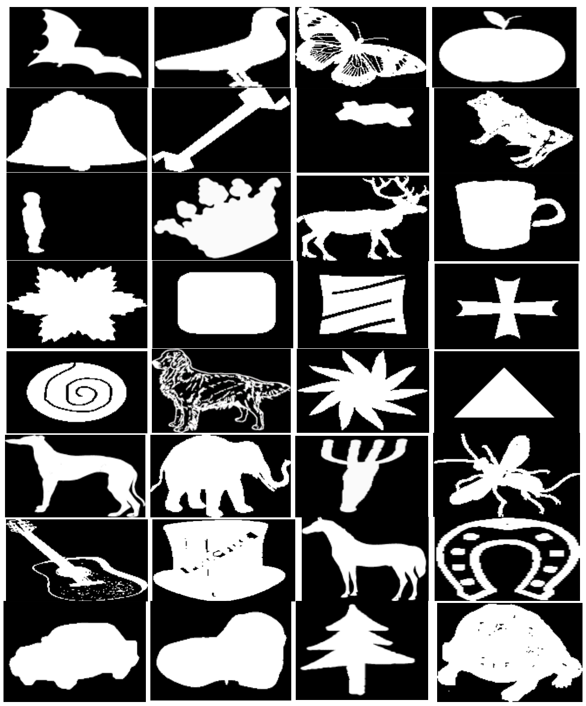


Fig. 6. Example of MPEG7 Database.

V. Results and Discussion

The problem treated in this work is not the spacing between samples, but simply the change in the order of points in storage or manipulation of those points. We notice that the Normal Fourier transform requires $O(N \log N)$ to compute N Fourier modes from N data points. Also, novel transform achieves the same $O(N \log N)$ computational complexity.

Tables I, II, III and IV we present the x and y coordinates of the points with different indices of the four shapes given in Fig. 1. The polor radius for each point is shown in Table V. In Tables VI, VII, VIII, and IX we present the normal Fourier transform for each shape given in Fig. 1, we can show no equality of coefficients. In Tables X, XI, XII, and XIII we present the novel Fourier transform for each shape. We see that the novel two components U and V are the same. It's clear from Tables VI, VII, VIII, IX, X, XI, XII, and XIII which the proposed Fourier transform is efficient.

TABLE. I. Coordinates X and Y for Shape (Fig.1.a)

Index	1	2	3	4	5	6
Coordinates x	1	1	1.5	2	2	1.5
Coordinates y	1	2	2	2	1	1

TABLE. II. Coordinates X and Y for Shape (Fig.1.b)

Index	1	2	3	4	5	6
Coordinates x	1	1.5	2	2	1.5	1
Coordinates y	1	1	1	2	2	2

TABLE. III. Coordinates X and Y for Shape (Fig.1.c)

Index	1	2	3	4	5	6
Coordinates x	1	1.5	2	2	1.5	1
Coordinates y	2	2	2	1	1	1

TABLE. IV. coordinates x and y for shape (Fig.1.d)

Index	1	2	3	4	5	6
Coordinates x	1	1	1.5	2	2	1.5
Coordinates y	2	1	2	1	2	1

TABLE. V. Radious of Shapes (Fig.1.a), (Fig.1.b), (Fig.1.c) and (Fig.1.d)

τ for shape 1.a)	1.41	2.23	2.50	2.82	2.23	1.80
τ for shape 1.b)	1.41	1.80	2.23	2.82	2.50	2.23
τ for shape 1.c)	2.23	2.50	2.82	2.23	1.80	1.41
τ for shape 1.d)	2.23	1.41	2.23	2.23	2.82	1.80

TABLE. VI. Normal Fourier Transform of Shape (Fig.1.a)

U	**V**
-0.0000 + 0.2887i	-0.3333 - 0.0000i
0.0000 + 0.0000i	0.0000 - 0.0000i
0 - 0.0000i	0.1667 - 0.0000i
-0.0000 - 0.0000i	-0.0000 - 0.0000i
-0.0000 - 0.2887i	-0.3333 - 0.0000i
1.5000 + 0.0000i	1.5000 + 0.0000i

TABLE. VII. Normal Fourier Transform of Shape (Fig.1.b)

U	**V**
-0.2500 + 0.1443i	0.1667 + 0.2887i
0.0000 - 0.0000i	0.0000 - 0.0000i
0.0000 - 0.0000i	0.1667 - 0.0000i
-0.0000 - 0.0000i	-0.0000 - 0.0000i
-0.2500 - 0.1443i	0.1667 - 0.2887i
1.5000 + 0.0000i	1.5000 + 0.0000i

TABLE. VIII. Normal Fourier Transform of Shape (Fig.1.c)

U	**V**
-0.2500 + 0.1443i	-0.1667 - 0.2887i
0.0000 - 0.0000i	0.0000 + 0.0000i
0.0000 - 0.0000i	-0.1667 - 0.0000i
-0.0000 - 0.0000i	-0.0000 - 0.0000i
-0.2500 - 0.1443i	-0.1667 + 0.2887i
1.5000 + 0.0000i	1.5000 + 0.0000i

TABLE. IX. NORMAL FOURIER TRANSFORM OF SHAPE (FIG.1.D)

U	*V*
-0.0000 + 0.2887i	0.0000 - 0.0000i
0.0000 + 0.0000i	0.0000 + 0.0000i
0 - 0.0000i	-0.5000 - 0.0000i
-0.0000 - 0.0000i	-0.0000 + 0.0000i
-0.0000 - 0.2887i	-0.0000 - 0.0000i
1.5000 + 0.0000i	1.5000 + 0.0000i

TABLE. X. NOVEL FOURIER TRANSFORM OF SHAPE (FIG.1.A)

U	V
-0.9558 - 0.9468i	-1.0020 - 0.9093i
0.0571 + 0.9518i	0.1659 + 0.9746i
0.2418 - 0.4530i	0.1739 - 0.5847i
-0.1428 + 0.2268i	-0.2092 + 0.3785i
0.1581 - 0.1539i	0.3136 - 0.1968i
-0.1102 - 0.0536i	-0.2207 - 0.1324i

TABLE. XI. NOVEL FOURIER TRANSFORM OF SHAPE (FIG.1.B)

U	V
-0.9558 - 0.9468i	-1.0020 - 0.9093i
0.0571 + 0.9518i	0.1659 + 0.9746i
0.2418 - 0.4530i	0.1739 - 0.5847i
-0.1428 + 0.2268i	-0.2092 + 0.3785i
0.1581 - 0.1539i	0.3136 - 0.1968i
-0.1102 - 0.0536i	-0.2207 - 0.1324i

TABLE. XII. NOVEL FOURIER TRANSFORM OF SHAPE (FIG.1.C)

U	V
-0.9558 - 0.9468i	-1.0020 - 0.9093i
0.0571 + 0.9518i	0.1659 + 0.9746i
0.2418 - 0.4530i	0.1739 - 0.5847i
-0.1428 + 0.2268i	-0.2092 + 0.3785i
0.1581 - 0.1539i	0.3136 - 0.1968i
-0.1102 - 0.0536i	-0.2207 - 0.1324i

TABLE. XIII. NOVEL FOURIER TRANSFORM OF SHAPE (FIG.1.D)

U	V
-0.9558 - 0.9468i	-1.0020 - 0.9093i
0.0571 + 0.9518i	0.1659 + 0.9746i
0.2418 - 0.4530i	0.1739 - 0.5847i
-0.1428 + 0.2268i	-0.2092 + 0.3785i
0.1581 - 0.1539i	0.3136 - 0.1968i
-0.1102 - 0.0536i	-0.2207 - 0.1324i

In Tables XIV and XV, we see that coefficients U and V are the same for two contours, only the first 10 coefficients using novel Fourier transform are presented. Graphic representation of $abs|\log(U)|$ and $abs|\log(V)|$ for all coefficients is given in Fig. 4. We note that coefficients are always in the same order of magnitude (see Table XVI).

TABLE. XIV. NOVEL FOURIER TRNASFORM OF COORDINATES X FOR TWO CONTOURS 4(A) AND 4(B)

U	U
1.5477 - 2.3974i	1.5477 - 2.3974i
-1.0556 - 2.2168i	-1.0556 - 2.2168i
-1.8839 - 0.1026i	-1.8839 - 0.1026i
-0.5257 + 1.1775i	-0.5257 + 1.1775i
0.7773 + 0.4266i	0.7773 + 0.4266i
0.3836 - 0.7878i	0.3836 - 0.7878i
-0.8005 - 0.6661i	-0.8005 - 0.6661i
-1.0159 + 0.4648i	-1.0159 + 0.4648i
-0.1105 + 1.0289i	-0.1105 + 1.0289i
0.6326 + 0.5173i	0.6326 + 0.5173i

TABLE. XV. NOVEL FOURIER TRANSFORM OF COORDINATES Y FOR TWO CONTOURS 4(A) AND 4(B)

V	V
1.5950 - 3.0604i	1.5950 - 3.0604i
-1.8277 - 2.4957i	-1.8277 - 2.4957i
-2.5307 + 0.5763i	-2.5307 + 0.5763i
-0.2078 + 2.0807i	-0.2078 + 2.0807i
1.6373 + 0.5444i	1.6373 + 0.5444i
0.7498 - 1.3746i	0.7498 - 1.3746i
-1.1345 - 1.0216i	-1.1345 - 1.0216i
-1.2702 + 0.7405i	-1.2702 + 0.7405i
0.2403 + 1.3063i	0.2403 + 1.3063i
1.0990 + 0.1632i	1.0990 + 0.1632i

TABLE. XVI. RESULT $abs|\log(U)|$ FOR FIG. 5(A) AND $abs|\log(U)|$ FOR FIG. 5(B)

U	U
5.6537	5.6538
5.5034	5.5035
5.2400	5.2399
4.8595	4.8591
4.4849	4.4839
4.4731	4.4717
4.6457	4.6445
4.7160	4.7150
4.6394	4.6388
4.4033	4.4035

VI. CONCLUSION

This paper presents a new Fourier transform to solve the problem of index of a point. The radius parameter is used in the development. The experimental results show that NUFT presents substantial advantages than normal Fourier transform. Shift theorem is available only where the Shift is linear, but when the order of point is randomly this theorem is not valid. The advantage of our transform is invariance under change of index point, and especially robustness to noise.

VII. Future Work

In the future, the authors are interested to implement our descriptor on speech signal and for 3D objects (mesh, surfaces) and using neural network.

References

[1] A. El-ghazal, O. Basir, and S. Belkasim, “Invariant curvature-based Fourier shape descriptors,” Journal of Visual Communication and Image Representation, 23(4), pp.622-633, 2012.

[2] Y. Xu, Z. Xie, Z. Chen, and L. Wu, “Shape similarity measurement model for holed polygons based on position graphs and Fourier descriptors,” International Journal of Geographical Information Science, 31(2), pp. 253-279, 2017.

[3] S.D. Choudhury, and T. Tjahjadi, “Silhouette-based gait recognition using Procrustes shape analysis and elliptic Fourier descriptors,” Pattern Recognition, 45(9), pp.3414-3426, 2012.

[4] C.Y. Wang, and Y. Zhang, “Affine invariant feature extraction algorithm based on multiscale autoconvolution combining with texture structure analysis,” Optik-International Journal for Light and Electron Optics, 125(13), pp. 3184-3189, 2013.

[5] P. Luo, X. Qu, X. Qing, and J. Gu, “CT Image Denoising Using Double Density Dual Tree Complex Wavelet with Modified Thresholding,” In 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), pp. 287-290, 2018.

[6] W. Zhao, and S. Du, “Learning multiscale and deep representations for classifying remotely sensed imagery,” ISPRS Journal of Photogrammetry and Remote Sensing, 113, pp. 155-165, 2016.

[7] S. Shi, R. Xiong, S. Ma, X. Fan, and W. Gao, “Image compressive sensing using overlapped block projection and reconstruction,” In 2015 IEEE International Symposium on Circuits and Systems (ISCAS), pp.1670-1673, 2015.

[8] D. Li, X. Zhang, and D. Tu, “ An FFT-Based Technique for Underwater Image Stitching,” In International Workshop of Advanced Manufacturing and Automation, pp. 527-535, 2018.

[9] D. Ravi, C. Wong, B. Lo, and G.Z. Yang, “Deep learning for human activity recognition: A resource efficient implementation on low-power devices, ” In 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN), pp. 71-76, 2016.

[10] A. Abdalla, and C. Green, “Utilisation of Fast Fourier Transform and Least-squares Modification of Stokes formula to compile a gravimetric geoid model over Khartoum State: Sudan,” Arabian Journal of Geosciences, 9(3), 236, 2016.

[11] Z. Li, H. Jia, Y. Zhang, T. Chen, L. Yuan, Cao L., and X. Wang, “AutoFFT: a template-based FFT codes auto-generation framework for ARM and X86 CPUs,” In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, p. 25, 2019.

[12] D. Takahashi, “Parallel FFT Algorithms for Shared-Memory Parallel Computers,” In Fast Fourier Transform Algorithms for Parallel Computers, pp. 69-76, 2019.

[13] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, “Simple and practical algorithm for sparse Fourier transform. In Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms,” Society for Industrial and Applied Mathematics, pp.1183-1194, 2012.

[14] A. Pfeifer, and V. Lohweg, “Detection of commercial offset printing using an adaptive software architecture for the DFT,” In 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1-4, 2016.

[15] S. Shi, R. Yang, and H. You, “A new two-dimensional Fourier transform algorithm based on image sparsity,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1373-1377, 2017.

[16] H. Li, X. Jin, N. Yang, and Z. Yang, “The recognition of landed aircrafts based on PCNN model and affine moment invariants,” Pattern Recognition Letters, 51, pp. 23-29, 2015.

[17] G., Plonka, D., Potts, G., Steidl, and M.Tasche, “Fast Fourier Transforms for Nonequispaced Data,” In Numerical Fourier Analysis, pp. 377-419, 2018.

[18] S.F., Potter, N.A., Gumerov, and R. Duraiswami, “Fast interpolation of bandlimited functions,” In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4516-4520, 2017.

[19] J. M. Lin, “Python Non-Uniform Fast Fourier Transform (PyNUFFT): multi-dimensional non-Cartesian image reconstruction package for heterogeneous platforms and applications to MRI,” arXiv preprint arXiv:1710.03197, 2017.

[20] R. Matthysen, and D. Huybrechs, “Function approximation on arbitrary domains using Fourier extension frames,” SIAM Journal on Numerical Analysis, 56(3), pp. 1360-1385, 2018.

[21] J. S. Walker, “Fast fourier transforms, ” CRC press, 2017.

[22] G. Plonka, D. Potts, G. Steidl, and M. Tasche, “Fast Fourier Transforms for Nonequispaced Data,” In Numerical Fourier Analysis, pp. 377-419, 2018.

[23] J. Schmid, T. Glatz, B. Zabihian, M Liu, W. Drexler, and O. Scherzer, “Nonequispaced grid sampling in photoacoustics with a nonuniform fast Fourier transform,” Journal of biomedical optics, 21(1), 015005, 2016.

[24] L. Le Magoarou, and R. Gribonval, “ Are there approximate fast fourier transforms on graphs,” In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4811-4815, 2016.

[25] D. Ruiz-Antolin, and A. Townsend, “A nonuniform fast Fourier transform based on low rank approximation,” SIAM Journal on Scientific Computing, 40(1), pp. A529-A547, 2018.

[26] T. Ou, “gNUFFTW: Auto-Tuning for High-Performance GPU-Accelerated Non-Uniform Fast Fourier Transforms, ” 2017.

[27] D. Potts, G. Steidl, and M. Tasche, “Fast Fourier transforms for nonequispaced data: A tutorial, ” InModern Sampling Theory: Mathematics and Application, Birkhäuser, Boston, MA,pp.247-270, 2001.

[28] L. Saribulut, T.E.K.E. Ahmet, and M. Tumay, “Fundamentals and literature review of Fourier transform in power quality issues,” Journal of Electrical and Electronics Engineering Research, 5(1) , pp. 9-22, 2013.

[29] de Jesus Romero-Troncoso, R. “Multirate signal processing to improve FFT-based analysis for detecting faults in induction motors, ” IEEE Transactions on industrial informatics, 13(3), pp. 1291-1300, 2016.

[30] M.B. Farah, R. Guesmi, A. Kachouri, and M. Samet, “A novel chaos based optical image encryption using fractional Fourier transform and DNA sequence operation,” Optics & Laser Technology, 121, 105777, 2020.

[31] Zhu, J., Balieu, R., Lu, X., and Kringos, N. (2018). “Microstructure evaluation of polymer-modified bitumen by image analysis using two-dimensional fast Fourier transform, ” *Materials & design*, *137*, 164-175.

[32] H.B. Kekre , T. Sarode, P. Natu, “ImageCompression Using Real Fourier Transform, Its Wavelet Transform And Hybrid Wavelet With DCT,” International Journal of Advanced Computer Science and Applications, 4(5), 2013.

[33] K. Aznag, N. Kane, A. El oirrak, and E. El bachari, “New approach for 3D object recognition using non uniform fourier and moments coefficients,” Journal of Engineering Science and Technology, 12(5) , pp.1354-1365, 2017.

[34] S. Bagchi, and S.K. Mitra, “The nonuniform discrete Fourier transform and its applications in signal processing,” Springer Science and Business Media, p. 463, 2010.

[35] X. Bai, X. Yang, L.J. Latecki, W. Liu, and Z. Tu, “Learning context-sensitive shape similarity by graph transduction,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(5), pp.861–874, 2010.

LEA-SIoT: Hardware Architecture of Lightweight Encryption Algorithm for Secure IoT on FPGA Platform

Bharathi R[1]
Research Scholar
PRIST University, Thanjavur
Tamil Nadu, India

N. Parvatham[2]
Associate Professor
PRIST University, Thanjavur
Tamil Nadu, India

***Abstract*—The Internet of Things (IoT) is one of the emerging technology in today's world to connect billions of electronic devices and providing the data security to these electronic devices while transmission from the attacks is a big challenging task. These electronic devices are smaller and consume less power. The conventional security algorithms are complex with its computations and not suitable for IoT environments. In this article, the hardware architecture of the new Lightweight encryption algorithm (LEA) for the secured Internet of things (SIoT) is designed, which includes Encryption, decryption along with key generation process. The New LEA-SIoT is a hybrid combination of the Feistel networks and Substitution-permutation Network (SPN). The encryption/decryption architecture is the composition of Logical operations, substitution transformations, and swapping. The encryption/decryption process is designed for 64-bit data input and 64-bit key inputs. The key generation process is designed with the help of KHAZAD block cipher algorithm. The encryption and key generation process are executing in parallel with pipelined architecture with five rounds to improve the hardware and computational complexity in IoT systems. TheLEA-SIoT is designed on the Xilinx platform and implemented on Artix-7 FPGA. The hardware constraints like area, power, and timing utilization are summarized. The Comparison of the LEA-SIoT with similar security algorithms is tabulated with improvements.**

***Keywords*—IOT Devices; Security algorithm; Encryption; Decryption; Key generation; FPGA**

I. Introduction

There are billions of devices connected on the internet, and a massive amount of data being generated from these devices and collect these data in IoT with authentication, service support, and privacy is a big challenging task. The future of the IoT is more in general to the public usage in-terms of smart homes, smart cities, virtual power plants, smart grids, intelligent transportation. To secure the data in the IoT environment, the lightweight security algorithms are suitable and essential because of less computational complexity. These lightweight algorithms are analyzed with a security level, hardware technology, throughput, latency, energy and power consumption, Memory utilization, and efficiency [1-3]. The IoT architecture is designed based on the layers used on the applications. The 3-layer architecture includes a physical, network, and application layer is used. The 4-layer architecture includes data perception, heterogeneous network access, data management, and intelligent service layer.5-layer IoT consists of Perception layer, network, processing layer, application, and business layer. The different attacks on IoT, which includes Denial of service (DoS), Man-in Middle, Wormhole, alteration, fabrication, and eavesdropping. The security challenges of IoT is security and data protection, authentication, privacy, access control, trust, and policy enforcement [4-6].

Section "A" discusses the background of the previous research works of Security algorithms for IoT applications followed by research gaps in section "B". Section II describes the Lightweight encryption algorithm (LEA) for the secured Internet of things (SIoT) with detailed hardware architecture. Section III explains the simulation results of the LEA-SIoT and analyze the hardware constraints of LEA-SIoT and comparison of the LEA-SIoT with other Security algorithms with improvements. Section IV concludes the overall work with improvements and future work.

A. The Background

This section discusses the existing work of different Security algorithms for IoT environments. The Goyal et al. [9] presents a Public key algorithm for low power IoT devices, which includes Elliptic Curve Diffie-Hellman (EC-DH) method. This EC-DH provides extended security with low power consumption for IoT gadgets with nominal processing speed and also compare with Diffie-Hellman (DH), and RSA algorithm with minor improvements. The Safi [10-11] describes the new hybrid encryption method to provide the security in IoT using public, private key along with digital signatures. The Advanced-encryption-standard (AES) is used for the public key, and NTRU is used for Private and digital signature. This hybrid method uses a software-based approach, which is not compactable with real-time hardware IoT environments. Khan et al. [12] present a performance analysis of different security algorithms like AES, SHA, and ECC using crypto++ library with C for small IoT applications and implemented on hardware Raspberry Pi-3. Landge et al. [13] describe the Message Digest (MD)-5 hashing algorithm for secure IoT using software platform. The sender and receiver hash information with execution time on server base is discussed. The Elliptic Curve scalar multiplication model is presented by Venugopal et al. [14] for IoT security with smaller key sizes. The multiplier has two variants of karatsuba to improve chip utilization. Hajj et al. [15] present the analysis of the cryptographic algorithms on a hardware platform for IoT devices. The algorithms include symmetric, asymmetric, and, hashing, along with signature algorithms, are demonstrated on the Raspberry –Pi model. The hybrid encryption algorithm is described by Chandu et al. [16] which includes the AES algorithm for data transmission in the cloud and RSA algorithm encrypts the AES key which is used by the authorized user. Only AES part is performed by FPGA hardware, and the RSA part is done by Matlab environments. Liang et al. [17] present an authentication algorithm for identification of data under the IoT environment. The Hausdorff distance (HS) based algorithm is used for position selection, characteristic matching, and identification of data. Guruprasad et al. [18] present an analysis of security algorithms on FPGA platform which includes AES with different key sizes, Data encryption standard (DES), Triple DES (TDES), and Light encryption device (LED). Samir et al. [19] explain the lightweight hardware security algorithms for IoT on ASIC and FPGA platform. The hardware complexity in the IoT environment is resolved using authentication block ciphers. The dynamically reconfigurable security algorithm is presented by Yao et al. [20] for IoT, which includes AES and TDES algorithms are used to realize the dynamic switching in the reconfigurable partition, and core controller controls it. Tao et al. [21] explain about hardware-based block cipher with secured data collection for IoT based Healthcare, and KATAN security algorithm is used to provide crypto information from the patients and to doctor via a secured share cloud server. Ahmed et al. [22] presents the lightweight encryption algorithm for secured IoT on the Matlab platform and also implemented on 8-bit Microcontroller.

B. *Research Gap*

From the review of recent literature, it has been noticed that the amount of work carried on Security algorithms for IoT environment is based on software approaches and few on hardware-based approaches like simple microcontroller and raspberry-Pi. In the available existing hardware-based approaches carried most of them with conventional security algorithms. These existing algorithms are facing hardware complexity, performance degradation, and more chip area consumption with massive power consumption in the IoT environment. The existing lightweight encryption algorithms are uses more computational rounds for the confusion, which leads to more chip area and affect the system performance. These lightweight encryption algorithms face significant security challenges in the IoT environment. Thus there is a need for "*cost-effective new lightweight encryption algorithm for Secure IoT environment.*".

II. Proposed Work

In this section, the proposed Lightweight encryption algorithm is designed for secure IoT (LEA-SIoT) is explained with its hardware architectures.

The proposed algorithm provides a suitable simple and efficient Hardware architecture for real-time IoT environment security. There are several block cipher algorithms with different network structures implemented with its own merits and demerits. In general, there are five different types of block ciphers available with different network structures namely Feistel network (FN), Substitution permutation networks (SPN), LFSR-based approach, Add-Rotate XOR (ARX) and hybrid combination approach.

In the proposed design, a hybrid approach is used with the combination of a Feistel network (FN) & SPN. The FN is having a major advantage of using almost similar encryption-decryption process. The SPN provides different rounds of substitution, which ensures the ciphertext is available in a pseudo-random manner. So the hybrid approach based architecture provides a new lightweight block cipher algorithm with efficient security and less computational complexity in IOT environment.

The lightweight encryption algorithm (LEA) for secured IoT (SIoT) is a symmetric key block-cipher (SKC) model which consist of 64-bit plain text and 64-bit key input. In general, most of the block cipher encryption process consists of many rounds of operations to keep the security process stronger. So each round of operation is designed based on some mathematics operations to create confusion. If the number of rounds increases, the system can provide better security with more resource utilization on a chip.

In the proposed design, only five rounds of operations are considered to improve the chip area and computational complexity in the IoT environment. The LEA-SIoT model mainly consists of encryption, decryption, and key generation process. Table I shows the notations with its functions used in the hardware architecture of the encryption, decryption, and key generation process. The detailed explanation of each process is discussed in the below section.

A. *Key Generation Module*

The complete data information, along with security is dependent on the key. If attackers know about the key, the data information will be lost. The encryption and decryption are processed with the same key input, which is having 64-bit. For each encryption-decryption round, the separate key generation input is provided to maintain the security. In this design, the Feistel-network based encryption algorithm is processed with five rounds using five different key inputs. The key generation provides the five unique keys to the encryption-decryption process.

TABLE. I. Notations

Symbol	Function
$\odot$	XNOR
$\oplus$	XOR
$\boxed{\text{R}}$	Register
$\boxed{\text{SF}}$	SBOX-Function
$\#$, #	Concatenation

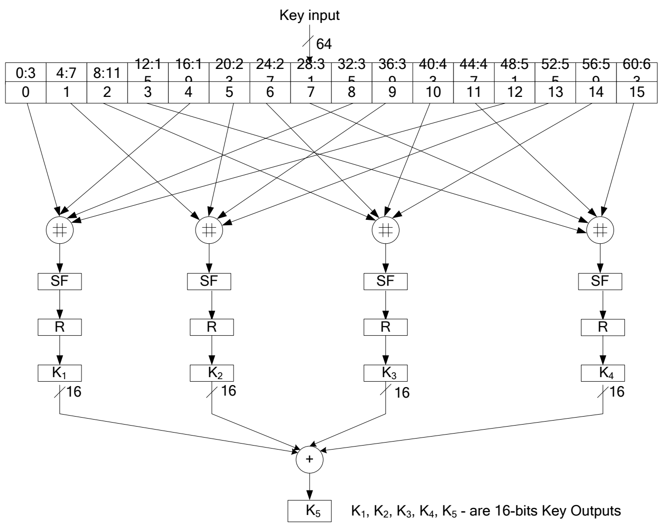


Fig. 1. Architecture of Key Generation Module.

The proposed algorithm provides a 64-bit key which is used to encrypt the 64-bit plain text data. The 64-bit key input (K_I) is considered from the user interest. The 64-bit key input is expanded in the key expansion block, and perform the substantial process to create the confusion for the given K_I to generate the five different keys. The key generation process mainly includes substantial process followed by concatenation and S-Box function (SF) to generate the five different keys like K_1, K_2, K_3, K_4, and K_5 are represented in Fig. 1. The key generation process is explained in the below steps as follows.

The 64-bit key input (K_I) is divided into 16 segments, each of 4-bits. Consider the 16-segments, perform initial substitution as per below equation (1).

$$K_{Si} = \sum_{j=1}^{4} K_{I\ 4(j-1)+i} \tag{1}$$

Where j = 1 to 4 for each block creation with 4-bits and i = 1 to 4 for the first four rounds of key process. The single 16-bit block is generated with 4-segments. The first block K_{S1} contains {KI_1, KI_7, KI_9, KI_{13}} as a first segment, {KI_2, KI_6, KI_{10}, KI_{14}} as a second segment, {KI_3, KI_7, KI_{11}, KI_{15}} as a third segment, and {KI_4, KI_8, KI_{12}, KI_{16}} as the fourth segment for first round key generation process. Similarly, for K_{S2}, $K_{S3}$and K_{S4}will be calculated for second, third, and fourth blocks in key round process, respectively.

The 16-bit four blocks K_{S1}, K_{S2}, K_{S3}, and K_{S4}, are passed to S-Box function (SF) as represented in the equation (2).

$$K_i = SF(K_{Si}) \tag{2}$$

The S-Box function (SF) mainly contains P and Q tables; these tables have linear and non-linear transformations, which results in confusion and diffusion are represented in Fig. 2. The S-Box Function is taken from the Khazad block cipher [7][8]. The linear and non-linear transformations are represented by P and Q table is tabulated in Table II and Table III, respectively. The four S-Box function (SF) outputs are stored in four temporary registers (R) and the register outputs represented as first four key outputs, namely, K1, K2, K3, and K4.

To generate the 5th key K5, perform the XOR operation of the first four round keys (K1, K2, K3, and K4), and it is represented in the below equation (3).

$$K_5 = \bigoplus_{i=1}^{4} K_i \tag{3}$$

B. Encryption Module

The encryption process has 64-bit plain text input along with 5-round keys input, and the Hardware architecture of the encryption process is represented in Fig. 3. The encryption process mainly composed with logical operators (XOR, XNOR), substitution transformations and swapping along with pipelined registers in the architecture. The encryption process runs parallel with a key generation process with round keys to improve the latency of the system.

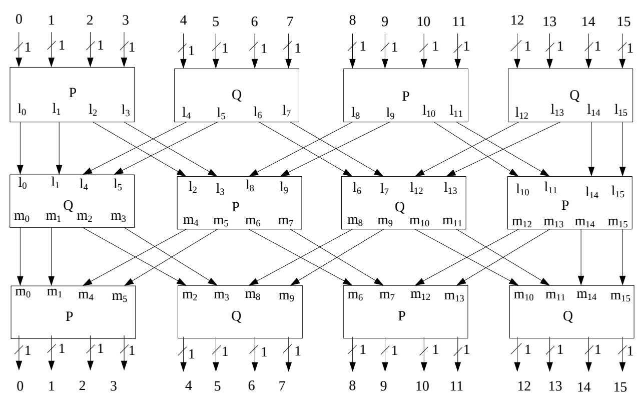


Fig. 2. Architecture of S-BOX Function (SF).

TABLE. II. P –TABLE FOR SBOX-FUNCTION

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
P(i)	3	F	E	0	5	4	B	C	D	A	9	6	7	8	2	1

TABLE. III. Q –TABLE FOR SBOX-FUNCTION

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Q(ii)	9	E	5	6	A	2	3	C	F	0	4	D	7	B	1	8

First, decompose the 64-bit plain text (P_T) into 4-blocks. Each block contains 16-bit input data. The plain text (P_T) has 4-blocks, P_T [0:15], P_T [16:31], P_T [32:47] and P_T [48:63]. These 4-blocks are used to process the encryption operation. These 4-blocks input data are processed in each round with swapping to altering the originality of bits, and it is necessary to create the confusion and diffusion about ciphertext. The corresponding round key Ki from key generation is performed the bitwise- XNOR operation with P_T [0:15] to generate the R_{11} and lk1 in the left-hand side. Similarly with P_T [48:63] to generate the R_{14} and rk1 in the right-hand side, respectively. The XNOR outputs lk1 and rk1 are feed separately to S-Box function (SF) to generate the lf1 and rf1 as represented in Fig. 3.

The same key generation S-Box-function (SF) is used in the encryption process. To perform the XOR operation of lf1 and rf1 separately with swapped plain text P_T [32:47] and P_T [16:31] to generate the R_{12} and R_{13} outputs respectively. The R_{11}, R_{12}, R_{13}, and R_{14} are first round encryption outputs and stored in a pipelined register (R) to maintain the synchronization with key generation process. In general, the round operation outputs are expressed in the below equation (4).

$$R_{i,j} = \begin{cases} P_{T\,i,j}\,\mathrm{XNOR}K_i\,;\ j = 1\,\&\,4 \\ P_{T\,i,j+1} \oplus lf_i\,;\ j = 2 \\ P_{T\,i,j-1} \oplus rf_i\,;\ j = 3 \end{cases} \tag{4}$$

Where i = 1 to 4 to perform the second, third, fourth, and fifth round transformation outputs. The final ciphertext will be generated after the 5th round transformation, and are shown in equation (5).

$$C_T = R_{51}\ \#\ R_{52}\ \#\ R_{53}\ \#\ R_{54} \tag{5}$$

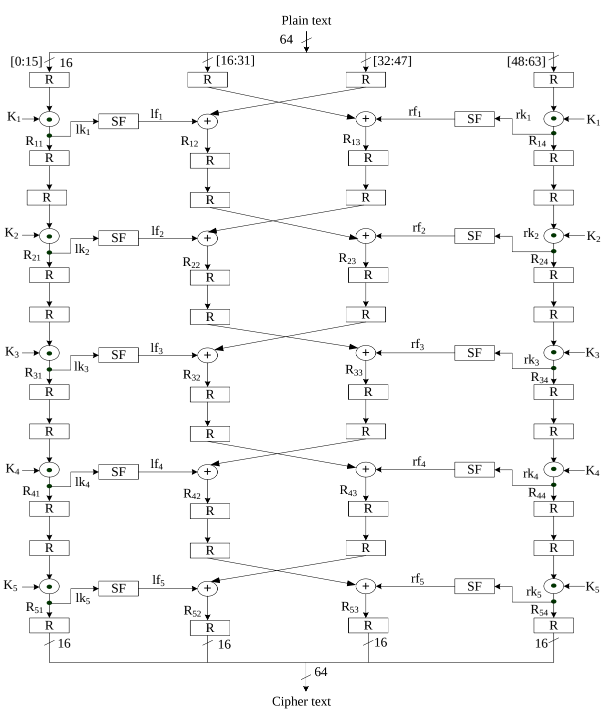


Fig. 3. Architecture of Encryption Module.

C. Decryption Module

The decryption process is almost similar to the encryption process with a few changes in the architecture, and it is represented in Fig. 4. The encryption process output 64-bit ciphertext as an input to the decryption process. First, decompose the 64-bit ciphertext (C_T) into 4-blocks. The 4-blocks are C_T [0:15], C_T [16:31], C_T [32:47] and C_T [48:63]. These 4-blocks are used to process the decryption operation. The corresponding round key Ki from key generation is performed the bitwise- XNOR operation with C_T [0:15] to generate the R_{11} and lk1 in the left-hand side. Similarly with C_T [48:63] to generate the R_{14} and rk1 in the right-hand side, respectively. The C_T [0:15] and C_T [48:63] are feed separately to S-Box function (SF) to generate the lf1 and rf1 as represented in Fig. 4.

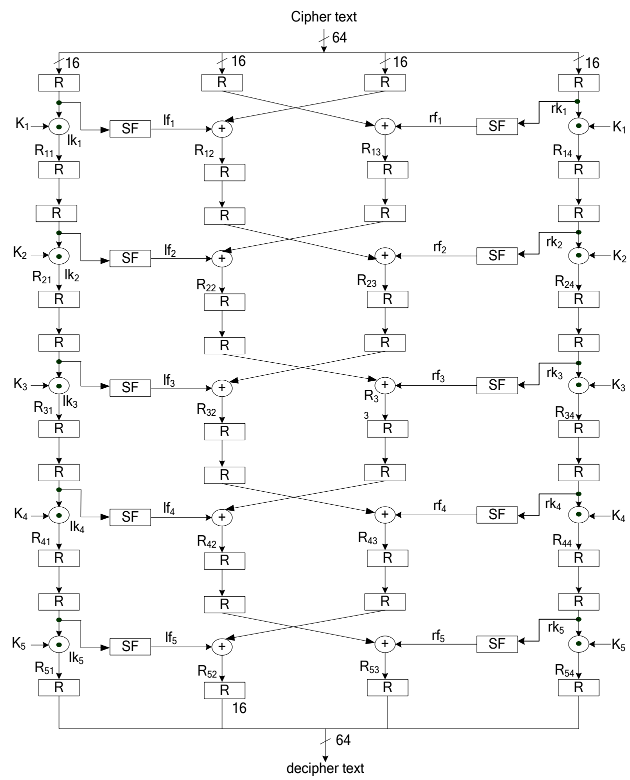


Fig. 4. Architecture of Decryption Module.

III. Results and Analysis

The proposed Lightweight encryption algorithm for secure IoT (LEA-SIoT) results are analyzed in the below section. The Proposed work is designed using Xilinx ISE 14.7 tool using Verilog-HDL language, and modelsim 6.5f is used for simulation. The proposed LEA-SIoT is prototyped and implemented on the Artix-7 FPGA Platform by considering Device-XC7A100T-3CSG324.

The LEA-SIoT simulation results are represented in Fig. 5. The global clock (clk) signal is activated with the positive edge along with active low reset (rst). Set the 64-bit data input (Plain_text) to 64’haaaa_bbbb_cccc_dddd and 64-bit key input (key_in) to 64’haaaa_aaaa_aaaa_aaaa. As per Encryption and decryption process, 64-bit Encryption output

(Cipher_text) 64'h5555_a1dd_1c2f_2222 and decryption output (decipher_text) 64'haaaa_bbbb_cccc_dddd will be generated. The decipher output is the same as the plain text input with the delay of 2 clock cycle. The encryption process is performed in parallel with 5 rounds and similarly, for the decryption process. This simulation result indicates that the lightweight encryption algorithm works effectively with low latency for IOT environment.

The integration of both Encryption and decryption are instantiated as a single module named as LEA-SIoT is synthesized and after a place and route operation, the hardware resource utilization in terms of area, time, power and speed are summarized in Table IV. The LEA-SIoT utilizes less chip area in terms of 211 slice registers, 1578 Slice LUT's and 200 LUT-FF pairs on FPGA. The LEA-SIoT operated a maximum frequency of 249.457 MHz with a minimum period of 4.009ns on Artix-7 FPGA. The power consumption report is generated using the X-Power analyzer tool with FPGA system frequency of 100 MHz. The LEA-SIoT consumes the 0.249W total power with the inclusion of 0.167W dynamic power.

The proposed LEA-SIoT (only Encryption) is compared concerning previous security algorithms like AES, RSA and 3DES [19] in terms of performance parameters like Area (Slice LUTs) and Power (mWatts) at 10 MHz frequency are tabulated in Table V. The performance parameter results are analyzed on the Zynq 7000 series FPGA with device XC7Z020 for all the security algorithms.

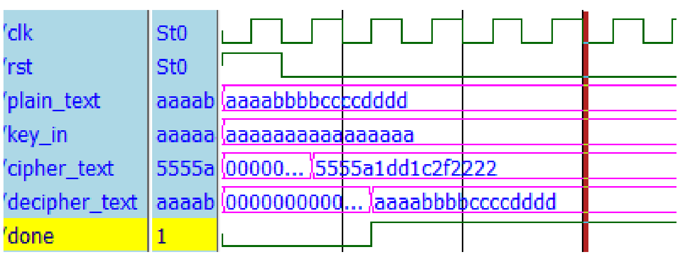


Fig. 5. Simulation Results of LEA-SIoT.

TABLE. IV. HARDWARE RESOURCE SUMMARY OF PROPOSED LEA (ENCRYPTION AND DECRYPTION) FOR SECURE IOT

Resources	**SIOT_ED**
Area	
Number of Slice Registers	211
Number of Slice LUTs	1578
Number of fully used LUT-FF pairs	200
Time	
Minimum period (ns)	4.009
Maximum Frequency (MHz)	249.457
Power	
Dynamic Power (W)	0.167
Total Power(W)	0.249
Speed	
Throughput (Gbps)	7.982

TABLE. V. SHOWS THE ENHANCEMENT OF AREA OVERHEAD (LUTS)

Design	**Frequency (MHz)**	**Area (LUTs)**	**Power (mW)**	**FPGA Device**
AES [19]	10	961	246	XC7Z020
RSA [19]	10	1178	255	XC7Z020
3DES [19]	10	1191	125	XC7Z020
Proposed Work	10	878	127	XC7Z020

The proposed design improves the area overhead (LUTs) around 10.49 % than AES, 25.56 % than RSA, and 26.28% than 3DES. The power consumption improved over around 48.37% than AES, and 50.19% than RSA at constant frequency 10 MHz (shown in Table V).

IV. CONCLUSION AND FUTURE WORK

The new Lightweight encryption algorithm (LEA) for the secured Internet of things (SIoT) is designed and implemented on Artix7 FPGA. The LEA-SIoT is a hybrid approach with Feistel networks and Substitution-permutation Network (SPN), for encryption/decryption.The LEA-SIoT reduces the computational and hardware complexity in IoT applications by executing the encryption/decryption and key generation process parallelly with pipeline architecture. The LEA-SIoT is simulated and synthesized on the Xilinx platform. The LEA-SIoT resource constraints like Area, time, Power, and speed utilizations are tabulated. The proposed algorithm for Secure IoT systems operate at 7.989 Gbps on Artix-7 FPGA and consume less power of 0.249W. The LEA-SIoT is compared with other security algorithms with improvements in the area (LUTs) and Power. The proposed LEA-SIoT (Encryption) model improves around 10.49% in Area and 48.37% in Power consumption than AES security algorithm. In the future, increase the number of encryption/decryption along with key generation rounds to strengthen the security with chip optimization, which suites in real time IoT applications.

REFERENCES

[1] Hatzivasilis, George, Konstantinos Fysarakis, Ioannis Papaefstathiou, and Charalampos Manifavas. "A review of lightweight block ciphers." Journal of Cryptographic Engineering 8, no. 2 (2018): 141-184.

[2] Surendran, Susha, Amira Nassef, and Babak D. Beheshti. "A survey of cryptographic algorithms for IoT devices." In 2018 IEEE Long Island Systems, Applications, and Technology Conference (LISAT), pp. 1-8. IEEE, 2018.

[3] Sadeeq, Mohammed AM, Subhi RM Zeebaree, Riyadh Qashi, Sarkar Hasan Ahmed, and Karwan Jacksi. "Internet of Things Security: A Survey." In 2018 International Conference on Advanced Science and Engineering (ICOASE), pp. 162-166. IEEE, 2018.

[4] Bhardwaj, Isha, Ajay Kumar, and Manu Bansal. "A review on lightweight cryptography algorithms for data security and authentication in IoTs." In 2017 4th International Conference on Signal Processing, Computing and Control (ISPCC), pp. 504-509. IEEE, 2017.

[5] Pawar, Ankush B., and Shashikant Ghumbre. "A survey on IoT applications, security challenges, and countermeasures." In 2016 International Conference on Computing, Analytics and Security Trends (CAST), pp. 294-299. IEEE, 2016.

[6] Naru, Effy Raja, Hemraj Saini, and Mukesh Sharma. "A recent review on lightweight cryptography in IoT." In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud)(I-SMAC), pp. 887-890. IEEE, 2017.

[7] Barreto, P. S. L. M., and Vincent Rijmen. "The Khazad legacy-level block cipher." Primitive submitted to NESSIE 97 (2000).

[8] Standaert, Francois-Xavier, Gael Rouvroy, Jean-Jacques Quisquater, and Jean-Didier Legat. "Efficient FPGA implementations of block ciphers KHAZAD and MISTY1." In Third NESSIE Workshop, pp. 6-7. 2002.

[9] Goyal, Tarun Kumar, and Vineet Sahula. "Lightweight security algorithm for low power IoT devices." In 2016 International Conference on Advances in Computing, Communications, and Informatics (ICACCI), pp. 1725-1729. IEEE, 2016.

[10] Safi, Amirhossein. "Improving the Security of the Internet of Things Using Encryption Algorithms." International Journal of Computer, Electrical, Automation, Control, and Information Engineering 11 (2017): 5.

[11] Yousefi, Afsoon, and Seyed Mahdi Jameii. "Improving the security of internet of things using encryption algorithms." In 2017 International Conference on IoT and Application (ICIOT), pp. 1-5. IEEE, 2017.

[12] Khan, Nuzhat, Nazmus Sakib, Ismot Jerin, Shaela Quader, and Amitabha Chakrabarty. "Performance analysis of security algorithms for IoT devices." In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp. 130-133. IEEE, 2017.

[13] Landge, Irfan A., and Hannan Satopay. "Secured IoT Through Hashing Using MD5." In 2018 Fourth International Conference on Advances in Electrical, Electronics, Information, Communication, and Bio-Informatics (AEEICB), pp. 1-5. IEEE, 2018.

[14] Venugopal, Ellappan, and Tadesse Hailu. "FPGA Based Architecture of Elliptic Curve Scalar Multiplication for IOT." In 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), pp. 178-182. IEEE, 2018.

[15] El-Haii, Mohammed, Maroun Chamoun, Ahmad Fadlallah, and Ahmed Serhrouchni. "Analysis of Cryptographic Algorithms on IoT Hardware platforms." In 2018 2nd Cyber Security in Networking Conference (CSNet), pp. 1-5. IEEE, 2018.

[16] Chandu, Y., KS Rakesh Kumar, Ninad Vivek Prabhukhanolkar, A. N. Anish and Sushma Rawal. "Design and implementation of hybrid encryption for the security of IOT data." In 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), pp. 1228-1231. IEEE, 2017.

[17] Huang, Yin, Wei Liang, Jing Long, Jianbo Xu, and Kuan-Ching Li. "A Novel Identity Authentication for FPGA Based IP Designs." In 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), pp. 1531-1536. IEEE, 2018.

[18] Guruprasad, S. P., and B. S. Chandrasekar. "An evaluation framework for security algorithms performance realization on FPGA." In 2018 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC), pp. 1-6. IEEE, 2018.

[19] Samir, Nagham. "ASIC and FPGA Comparative Study for IoT Lightweight Hardware Security Algorithms." Ph.D. diss., Faculty of Engineering, Cairo University, Giza, 12613, Egypt, 2018.

[20] Wang, Zhu, Yan Yao, Xiaojun Tong, Qinghua Luo, and Xiangyu Chen. "Dynamically Reconfigurable Encryption and Decryption System Design for the Internet of Things Information Security." Sensors 19, no. 1 (2019): 143.

[21] Tao, Hai, Md Zakirul Alam Bhuiyan, Ahmed N. Abdalla, Mohammad Mehedi Hassan, Jasni Mohamad Zain, and Thaier Hayajneh. "Secured data collection with hardware-based ciphers for iot-based healthcare." IEEE Internet of Things Journal 6, no. 1 (2019): 410-420.

[22] Usman, Muhammad, Irfan Ahmed, M. Imran Aslam, Shujaat Khan, and Usman Ali Shah. "SIT: a lightweight encryption algorithm for secure internet of things." arXiv preprint arXiv: 1704.08688 (2017).

Delay-Aware and User-Adaptive Offloading of Computation-Intensive Applications with Per-Task Delay in Mobile Edge Computing Networks

Tarik Chanyour*[1], Youssef HMIMZ[2], Mohamed EL GHMARY[3], Mohammed Ouçamah CHERKAOUI MALKI[4]
FSDM, LIIAN
Sidi Mohamed Ben Abdellah University
P.O. Box 1796, Atlas-Fez, Morocco

***Abstract*—Mobile-edge computing (MEC) is a new paradigm with a great potential to extend mobile users capabilities because-of its proximity. It can contribute efficiently to optimize the energy consumption to preserve privacy, and reduce the bottlenecks of the network traffic. In addition, intensive-computation offloading is an active research area that can lessen latencies and energy consumption. Nevertheless, within multi-user networks with a multi-task scenario, select the tasks to offload is complex and critical. Actually, these selections and the resources' allocation have to be carefully considered as they affect the resulting energies and delays. In this work, we study a scenario considering a user-adaptive offloading where each user runs a list of heavy computation-tasks. Every task has to be processed in its associated MEC server within a fixed deadline. Hence, the proposed optimization problem target the minimization of a weighted-sum normalized function depending on three metrics. The first is energy consumption, the second is the total processing delays, and the third is the unsatisfied processing workload. The solution of the general problem is obtained using the solutions of two sub-problems. Also, all solutions are evaluated using a set of simulation experiments. Finally, the execution times are very encouraging for moderate sizes, and the proposed heuristic solutions give satisfactory results in terms of users cost function in pseudo-polynomial times.**

***Keywords*—Mobile edge computing; user-adaptive offloading; computation-intensive offloading; per-task delay; tasks satisfaction optimization**

I. Introduction

A variety of recent Smart Mobile Devices (SMD) have capabilities to process some emerging attractive computation-intensive applications. Besides, these applications which are resource-hungry led to appear a novel kind of constraints related to resources insufficiency. They yield new defies particularly with regard to energetic and latency concerns. The offloading technique [1] in the context of Mobile Edge Computing (MEC) [2], [3] offers a chance to free these devices from generated heavy tasks by migrating them to a nearby edge infrastructure with extra powerful resources. Therefore, the power consumption of a mobile device as well as the latency responses can be reduced if the device can offload some of its heavy tasks to a MEC server.

The multi-user offloading problems with the single-task scenario in MEC networks was extensively considered. Previous works studied the energy consumption or processing delay optimization. To reduce the overall energy consumption, the authors of [4], studied the offloading decisions and the allocation of communication resources. Alike, in [5] an optimization problem is derived to select the best offloading policy while saving the energy consumption. The next work [6] present, in our knowledge, the first try to enhance the energy consumption while considering devices with multiple independent tasks. But, the authors impractically consider mobile devices with the same tasks' number. Lately, the authors of [7] studied a single-user scenario with multi-task setting while they optimize radio resources and local frequency. For the issues of joint resources optimization in MEC networks, many previous works jointly addressed the radio and computation resources optimization for the multi-user scenarios [8]. Besides, in [9] the authors jointly decide the allocation of resources (remote computation and communication) to optimize the energy consumption while devices intend to offload their tasks to a MEC server within a 5G heterogeneous network. Likewise in [10], the authors jointly optimize the offloading decisions and resources' allocation (both computation and communication). Nevertheless, they neglected the energy consumption during tasks processing at the server. Also, withing a cloud Fog environment in [11], the problem of resource allocation to optimize energy consumption with load balancing is studied.

Different to [12] where the system's energy only is optimized while all the offloadable tasks of a given SMD are constrained to the same delay, this paper presents a generalization scenario regarding mainly the following three points: the first generalization point relies on the per-task delay property where we consider every task with its proper delay constraint. The second point concerns the existence of a one-to-many association that links every SMD to many available MEC servers. Accordingly, each task has to be processed in its associated ES. Finally, the third point relies on the user-adaptive offloading property where we target the optimization of a normalized objective function that considers not only the energy consumption metric, but two other important metrics. The first is the sum of the accumulated processing delays, and the second is the total offloaded workload. Thus, the proposed system architecture offers the possibility for every user to adapt its offloading concerns according to its needs or/and constraints.

We organized the rest of this work as follows. We describe the system's model in Section II. The obtained optimization problems are presented in Section III, and their resolution approach is summarized in Section IV. Evaluation and results

Fig. 1. A multi-server multi-task mobile edge-computing system

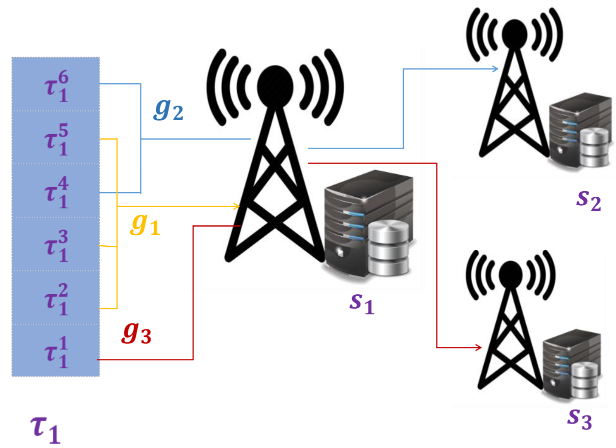


Fig. 2. Group-server association example

are presented in Section V. Finally, Section VI concludes the paper.

II. System Model

The adopted system's architecture in this work is shown in Fig. 1. It involves a set of Macro Cell (MC) where each MC is optionally equipped with an Edge Server (ES). The set of M available, supposed heterogeneous, edge servers is denoted $\mathbb{S} = \{s_1, s_2, ..., s_M\}$. Besides their traditional services, all macro-cells with ES can provide a set of independent computation-offloading services using the virtualization technology. Furthermore, we propose a distributed solution in each MC that considers its communication resources and the offloading decisions related to all SMDs within its coverage. Accordingly, a given MC, in this work, consists of a set of N Smart Mobile Devices (SMDs) denoted $\mathbb{E} = \{e_1, e_2, ..., e_N\}$. Each SMD is characterized by the parameters that are briefly presented in Table I.

Mainly, SMD i holds a list of n_i independent heavy tasks denoted $\tau_i = \{\tau_i^1, \tau_i^2, ..., \tau_i^{n_i}\}$. These tasks are assumed to be computationally intensive and delay sensitive. In addition, each task is viewed as an atomic input-data task, and cannot be divided into sub-tasks. Moreover, it is characterized by the following three properties $\tau_i^j \triangleq \langle d_i^j, \lambda_i^j, L_i^j \rangle$. In bits, the first property refers to the amount of the input parameters and program codes to transfer from the SMD to the edge server. In cycles, the second property specifies the workload referring to the computation amount needed to accomplish the task's processing. In seconds, the third property identifies the maximum tolerated delay within which τ_i^j has to be processed locally or at its associated edge server.

Moreover, within a given SMD i, its tasks are grouped into a set of n_i^g groups denoted $\mathbb{G}_i = \{g_1^1, g_1^2, ..., g_1^{n_i^g}\}$. Each group g_i^j is associated to a service hosted in an ES. The tasks of a given group are processed locally or transmitted via the current MC to their associated edge server. Accordingly, the processing frequencies related to SMD i are f_i^L for its local processing, and $f_{i,k}^S$ for remote processing at server s_k. For ease of use, the edge server s_k is denoted k. Moreover, we use a mapping function $g_i(.)$ to represent the association GROUP-SERVER of the tasks' of SMD i. Fig. 2 shows the list τ_1 and the corresponding group-server associations. Accordingly, the mapping function $g_1(.)$ is defined by the following mapping list $[(j, g_1(j))]_{j \in [\![1;6]\!]}$. This list contains three groups where the first $g_1 = \{\tau_1^2, \tau_1^3, \tau_1^5\}$ is associated to server s_1; the second group $g_2 = \{\tau_1^4, \tau_1^6\}$ is associated to server s_2, and the third $g_3 = \{\tau_1^1\}$ is associated to server s_3. Thus, the g_1 function is given by the mapping list $[(2,1),(3,1),(5,1),(4,2),(6,2),(1,3)]$. Here $g_1(j) = k$ refers to an association of task τ_1^j with server s_k. Moreover, it is assumed that the offloading process to remote servers experiences additional delay. Similar to the model in [7], we consider that this delay is proportional to the length of the data with a scaling factor δ_k that depends on the backhaul between the remote MC that shelter the ES s_k and the current MC. Then, the total experienced delay of a d bytes task equals $\delta_k.d$. Moreover, due to the complexity of evaluating the delay and energy consumption of the communication process between remote MCs and without loss of generality, we ignore the power consumption occurred at the backhaul network.
Further, the network model of our previous work [12] is adopted. Thus, we assume that e_i uses an estimated uplink rate r_i in each allocated subchannel. Additionally, with a number β_i of allocated subchannels, the total of its allocated uplink rate is $R_i = \beta_i r_i$.

A. The Offloading Model

All tasks are time constrained and can be processed either locally or offloaded to the edge server. Thus, the binary offloading decision variable for task τ_i^j is denoted α_i^j where $\alpha_i^j = 1$ refers to an offloading decision, whereas $\alpha_i^j = 0$ indicates a local processing decision.

$$\alpha_i^j \in \{0;1\} \qquad ; i \in \mathbb{E}; j \in [\![1; n_i]\!] \tag{1}$$

Additionally, the processing satisfaction of task τ_i^j is introduced using the binary variable γ_i^j where $\gamma_i^j = 0$ refers to a

TABLE I. IMPORTANT NOTATIONS

Notation	Definition
$\mathbb{E}$	Set of smart mobile devices
N	Total number of smart mobile devices
K	Total number of allocatable subchannels
K_t	Maximum allocatable number of subchannels per-SMD
N_t	Number of tasks threshold to enable "Heuristic Tasks Distribution"
N_m	Average tasks' number for all SMDs
n_i	Number of tasks handled by SMD i
τ_i	Computation tasks' set of SMD i
π_i	The parameters' set of SMD i
$D_{i,0}, D_i$	Total initial and offloaded tasks' data size at SMD i
$\Lambda_{i,0}, \Lambda_i$	Total initial and offloaded tasks' workload at SMD i
f_i^L	Local CPU frequency of SMD i (cycles/s)
$f_{i,k}^S$	Allocated CPU frequency for SMD i at s_k (cycles/s)
ξ_i^L	Energy coefficient depending on the chip architecture of SMD i and f_i^L
$\xi_{i,k}^S$	Energy coefficient depending on the chip architecture of s_k and $f_{i,k}^S$
p_i^T	Data transmission power of SMD i
r_i	Uplink rate of SMD i
$\delta_{i,k}$	Backhaul delay scaling factor related to s_k in SMD i

satisfied processing, otherwise $\gamma_i^j = 1$.

$$\gamma_i^j \in \{0;1\} \qquad ; i \in \mathbb{E}; j \in [\![1; n_i]\!] \tag{2}$$

This variable is defined such that $t_{i,j}^L + t_{i,j}^O \leqslant L_i^j \iff \gamma_i^j = 0$. Here L_i^j is the τ_i^j task's latency requirement. $t_{i,j}^L + t_{i,j}^O$ is the time to process τ_i^j and it is given in equations (7) and (10). This expression is formulated using the following constraint where M is a sufficiently large constant:

$$-M\gamma_i^j \leqslant L_i^j - t_{i,j}^L - t_{i,j}^O < M(1-\gamma_i^j) \qquad ; i \in \mathbb{E}; j \in [[1; n_i]] \tag{3}$$

The total number of subchannels assigned to e_i is denoted β_i where:

$$\beta_i \in [\![0; K_t]\!] \qquad ; i \in \mathbb{E} \tag{4}$$

Here K_t is a threshold value that is chosen according to N in such a way that it can take K where $N = 1$; and decreases with increasing value of N. Additionally, the sum of all allocated subchannels must not exceed K the total available subchannels, which gives:

$$\sum_{i \in \mathbb{E}} \beta_i \leqslant K \tag{5}$$

The decision $\beta_i = 0$ forbids offloading for e_i. In this case, it has to locally process all its tasks; whereas $\beta_i \neq 0$ indicates that e_i has to offload at least one task. This fact leads to the following offloading property $(\beta_i = 0) \iff (\sum_{j=1}^{n_i} \alpha_i^j = 0)$ which must hold for every SMD i. It can be formulated as:

$$\beta_i + \sum_{j=1}^{n_i} \alpha_i^j \leqslant 2\beta_i \sum_{j=1}^{n_i} \alpha_i^j \qquad ; i \in \mathbb{E} \tag{6}$$

B. Processing Model

If e_i locally executes task τ_i^j, its processing time in seconds lasts $\frac{\lambda_i^j}{f_{i,L}}$. Then, with its processing order, the actual local processing time of task τ_i^j is:

$$t_{i,j}^L = (1-\alpha_i^j) \sum_{k=1}^{j} (1-\alpha_i^k) \frac{\lambda_i^k}{f_{i,L}} \tag{7}$$

If e_i offload task τ_i^j to server s_k, its offloading time $t_{i,j}^O$ includes both, the transmission time $t_{i,j}^{Trans}$ and the waiting time $t_{i,j}^{Wait}$; which gives $t_{i,j}^O = t_{i,j}^{Trans} + t_{i,j}^{Wait}$. The second part includes the backhaul delay $t_{i,j}^{Delay}$, the execution time $t_{i,j}^{Exec}$, and the time to receive the result out from the server $t_{i,j}^{Res}$; which gives $t_{i,j}^{Wait} = t_{i,j}^{Delay} + t_{i,j}^{Exec} + t_{i,j}^{Res}$. Because the data size of the result is much smaller than the input data size, we ignore the time and the energy consumption of receiving out the result(see [6], [8], [13], [14]). Furthermore, without considering the delays relative to the processing order of task τ_i^j, these delays in seconds are given by :

$$\left(t_{i,j}^{Trans}, t_{i,j}^{Delay}, t_{i,j}^{Exec}\right) = \left(\frac{d_i^j}{\beta_i r_i}, \delta_{g_i(j)} d_i^j, \frac{\lambda_i^j}{f_{i,g_i(j)}^S}\right) \tag{8}$$

Now, with its processing order, the actual transmission delay of task τ_i^j is $t_{i,j}^{Trans} = \frac{\alpha_i^j}{\beta_i r_i} \sum_{k=1}^{j} \alpha_i^k d_i^k$. For e_i and with the sequential processing of the tasks' group in each server, the waiting delay of task τ_i^j is:

$$t_{i,j}^{Wait} = \delta_{g_i(j)} \sum_{k=1}^{j} \mathbb{1}(g_i(j)=g_i(k)) \alpha_i^k d_i^k + \frac{1}{f_{i,g_i(j)}^S} \sum_{k=1}^{j} \mathbb{1}(g_i(j)=g_i(k)) \alpha_i^k \lambda_i^k \tag{9}$$

Subsequently, its offloading delay is:

$$t_{i,j}^O = \frac{\alpha_i^j}{\beta_i r_i} \sum_{k=1}^{j} \alpha_i^k d_i^k + \sum_{k=1}^{j} \left(\mathbb{1}(g_i(j)=g_i(k)) \alpha_i^k \left(\delta_{g_i(j)} d_i^k + \frac{\lambda_i^k}{f_{i,g_i(j)}^S}\right)\right) \tag{10}$$

Also, for e_i and the tasks' processing orders we have:

$$t_i^{Trans} = \frac{1}{\beta_i r_i} \sum_{j=1}^{n_i} \alpha_i^j d_i^j = \frac{D_i}{\beta_i r_i} \tag{11}$$

C. Energetic Model

In this section, all energy expressions are given in Joule. Then, the processing of task τ_i^j consumes the amount of energy $e_{i,j} = \xi_i \lambda_i^j$ where ξ_i is a power coefficient depending on the processing unit's chip architecture and the processing frequency [9], [14]. Thus, the local energy consumption of task τ_i^j is $e_{i,j}^L = \xi_{i,L} \lambda_i^j$. Accordingly, the e_i's local energy consumption is:

$$e_i^L = \xi_{i,L} \sum_{j=1}^{n_i} \left(1-\alpha_i^j\right) \lambda_i^j = \xi_{i,L}(\Lambda_{i,0} - \Lambda_i) \tag{12}$$

Here, $\xi_{i,L}\Lambda_{i,0} = E_{i,0}^L$ is the energy consumption of the local processing of all tasks.

Furthermore, the energy consumption of the offloading process is $e_i^O = e_i^{Trans} + e_i^{Exec}$. Here e_i^{Trans} is the transmission consumption occurred at e_i, and e_i^{Exec} is the energy consumption occurred during the processing of offloaded tasks within all ESs. The first is obtained by multiplying the transmission period in the offloading processes by the transmission power. Thus:

$$e_i^{Trans} = P_i^T t_i^{Trans} = \frac{P_i^T D_i}{\beta_i r_i} \tag{13}$$

Again, the energy consumption while executing task τ_i^j at its dedicated edge server is $e_{i,j}^{Exec} = \xi_{i,g_i(j)}^S \lambda_i^j$ [9], [14]. As a result, the overall offloading energy consumption occurred at e_i is :

$$e_i^O = \begin{cases} \dfrac{P_i^T D_i}{\beta_i r_i} + \sum_{j=1}^{n_i} \xi_{i,g_i(j)}^S \alpha_i^j \lambda_i^j & ;\beta_i \neq 0 \\ 0 & ;\beta_i = 0 \end{cases} \tag{14}$$

D. The Cost Function

Now, the total e_i's energy consumption can be formulated using its tasks' offloading allocations given by a vector of n_i binary variables: $\alpha_i = \left(\alpha_i^1, ..., \alpha_i^{n_i}\right)$ and its allocated number of uplink subchannel(s) β_i. This energy includes the local processing consumption (if some tasks are locally processed) and the offloading process consumption (if some tasks are offloaded). It is formulated as:

$$E_i(\alpha_i,\beta_i)=\begin{cases} \xi_{i,L}(\Lambda_{i,0}-\Lambda_i)+\dfrac{P_i^T D_i}{\beta_i r_i}+\sum\limits_{j=1}^{n_i}\xi_{i,g_i(j)}^S\alpha_i^j\lambda_i^j & ;\beta_i\neq 0 \\ \xi_{i,L}\Lambda_{i,0} & ;\beta_i=0 \end{cases} \tag{15}$$

Similarly, w.r.t. α_i and β_i, the e_i's overall processing time denoted $T_i\left(\alpha_i, \beta_i\right)$ includes the processing delays of the local and the offloaded tasks using equations (7) and (10). Thus, we have:

$$T_i\left(\alpha_i, \beta_i\right) = \sum_{j=1}^{n_i} \left(t_{i,j}^L + t_{i,j}^O\right) \tag{16}$$

Finally, w.r.t. the e_i's tasks satisfactions given by a n_i binary variables' vector $\gamma_i = \left(\gamma_i^1, ..., \gamma_i^{n_i}\right)$, the total workload of its unsatisfied tasks is:

$$W_i\left(\gamma_i\right) = \sum_{j=1}^{n_i} \gamma_i^j \lambda_i^j \qquad ;i \in \mathbb{E} \tag{17}$$

At this stage, we formulate a multi-criteria offloading with an elastic budget model for every SMD. Thus, each device looks at three metrics: its energy consumption and latencies while processing all its tasks, and its unsatisfied total processing workload. The proposed multi-objective function for each SMD is expressed with a weighted sum of the three proposed metrics. Moreover, the contribution of e_i is formulated using the following weighted-sum function:

$$\mathbb{F}_i(\alpha_i,\beta_i,\gamma_i)=x_i\frac{E_i(\alpha_i,\beta_i)}{E_{i,0}^L}+y_i\frac{T_i(\alpha_i,\beta_i)}{L_i}+(1-x_i-y_i)\frac{W_i(\gamma_i)}{\Lambda_{i,0}} \tag{18}$$

Here, x_i and y_i are two parameters related to SMD i that determine its offloading policy. Their values are choosen by the SMD policy such that the following three weights x_i, y_i and $1 - x_i - y_i$ are in the interval [0,1]. By deciding these weights, the user can adjust the priority to give to each metric. This operation depends on its preference regarding delay-sensitivity, energetic constraints, and its processing capability. For example, (x_i, y_i) can be set to (0,1) for SMDs running delay sensitive applications whereas they can be set to (1,0) for energy-constrained devices. Also, for SMDs with bad processing capability, (x_i, y_i) can be set to (0,0).
Additionally, the three denominators in this expression are variable-independent terms that serve to normalize the cost function. As a result, the overall cost function of all SMDs is finally expressed as:

$$\mathbb{F}\left(\alpha, \beta, \gamma\right) = \sum_{i\in\mathbb{E}} \mathbb{F}_i\left(\alpha_i, \beta_i, \gamma_i\right) \tag{19}$$

The variables are given by α (a global vector composed of N vectors, each vector α_i of length n_i contains the e_i's tasks offloading decisions), γ (a global vector composed of N vectors, each vector γ_i of length n_i contains the e_i's tasks satisfactions variables) and β (a global vector containing N radio spectrum allocation β_i for all SMDs).

III. The Optimization Problems

A. The General Problem Formulation

In our proposed optimization problem, we target to minimize three important metrics that influence the system's performance and the users' satisfaction. Accordingly, we formulate our optimization problem which we denote problem $\mathscr{P}1$ as :

$$\begin{aligned} &\mathscr{P}1\text{:}\underset{\{\alpha,\beta,\gamma\}}{\text{minimize}}\mathbb{F}(\alpha,\beta,\gamma) \\ &\text{s.t. } (1),(2),(3),(4),(5),(6) \end{aligned} \tag{20}$$

B. Problem Decomposition

The $\mathscr{P}1$ problem is an integer programming optimization problem. To deal with its high computational complexity, we decompose it into a sub-optimal scheme. $\mathscr{P}1$ decides for the offloading and satisfaction of tasks as well as the subchannels allocations. Thus, we derive a set of sub-problems denoted User-adaptive Offloading and Satisfaction Decisions (UOSD) where each instance concerns a given SMD and a subchannels allocation.

1) The User-Adaptive Offloading and Satisfaction Decisions sub-problem: The UOSD sub-problem for user e_i is denoted $\mathscr{P}2(i, \beta_i)$. Its formulation uses a known fixed number of subchannels β_i which satisfies $\beta_i \in [\![1; K_t]\!]$. Then, the objective function is $\mathbb{F}_i\left(\alpha_i, \beta_i, \gamma_i\right)$. Accordingly, the variables of decision are the following two vectors $\alpha_i = \left(\alpha_i^1, \alpha_i^2, ..., \alpha_i^{n_i}\right)$ and $\gamma_i = \left(\gamma_i^1, \gamma_i^2, ..., \gamma_i^{n_i}\right)$. Finally, it is formulated as:

$$\begin{array}{lll} \mathscr{P}2(i,\beta_i)\text{:} & \underset{\{\alpha_i,\gamma_i\}}{\text{minimize}} \quad \mathbb{F}_i(\alpha_i,\beta_i,\gamma_i) & \\ \text{s.t. } (C_{21}) & \alpha_i^j,\gamma_i^j\in\{0;1\} & ;j\in[\![1;n_i]\!]. \\ (C_{22}) & -M\gamma_i^j\leqslant L_i^j-t_{i,j}^L-t_{i,j}^O<M(1-\gamma_i^j) & ;j\in[\![1;n_i]\!]. \\ (C_{23}) & \sum\limits_{j=1}^{n_i}\alpha_i^j\geqslant 1 & \end{array} \tag{21}$$

2) The Subchannels Allocations sub-problem: The UOSD sub-problem's instances related to e_i are obtained by varying the value of β_i in the interval $[\![1; K_t]\!]$ with only one possible allocation. The expected result is the vector of subchannels allocation given by $\beta = (\beta_1, \beta_2, ..., \beta_N)$. Consequently, we build a matrix M with N rows and $K_t + 1$ columns. The first column of M contains the cost functions obtained with $\beta_i = 0$. For the other K_t columns, every cell M_{ij} stores the solution of problem $\mathscr{P}2(i, j)$. Then, to decide an allocation of β_i in row M_i, we use a binary indicator x_i^j to denote the decision of the M_{ij} element selection in row M_i ($x_i^j = 1$ refers to the allocation $\beta_i = j$ and $x_i^j = 0$ leads to ignore the M_{ij} cell in the cost function). Furthermore, one possible value for β_i gives the constraint $\sum_{j=0}^{K_t} x_i^j = 1$. Additionally, the total allocation of subchannels for e_i in row i is exactly $\sum_{j=0}^{K_t} j.x_i^j$. Therefore, the total allocation of subchannels gives the number $\sum_{i\in\mathbb{E}}\sum_{j=0}^{K_t} j.x_i^j$ and must be less or equal to K. Consequently, the general formulation of the resulting problem becomes:

$$\begin{aligned}
&\mathscr{P}3: && \underset{\{x_i^j\}}{\text{minimize}} \sum_{i\in\mathbb{E}}\sum_{j=0}^{K_t} x_i^j M_{ij} \\
&\text{s.t. } (C_{31}) && \sum_{i\in\mathbb{E}}\sum_{j=0}^{K_t} j.x_i^j \leqslant K. \\
&(C_{32}) && \sum_{j=0}^{K_t} x_i^j = 1 \qquad ; i\in\mathbb{E}. \\
&(C_{33}) && x_i^j \in \{0;1\} \qquad ; i\in\mathbb{E}; j\in[\![0;K_t]\!].
\end{aligned} \tag{22}$$

This formulation corresponds to a general case of the LSP problem presented in [15]. Also, it is the minimization form of a special case of the Multiple-Choice Knapsack Problem (MCKP) [16] where the weights are the integers in $[\![0;K_t]\!]$. For more details about this last problem, interested readers can refer to [17], [16]

IV. Problems' Resolution

To solve problem $\mathscr{P}3$ we propose two algorithms: the first is the recent BISSA [17] solution which demonstrated its effectiveness in our previous work [12]. It is a pseudo-polynomial time complexity algorithm. The second is the Exact Brute Force Search algorithm which we denote (MCKP-BFS). With the condition $K \gg K_t$, its time complexity equals $O({K_t}^N)$. Its pseudo-code is summarized in Algorithm 1. To run this algorithm, we use the matrix M that is build s.t. its first column contains the results of the local processing. Its last K_t columns of N rows are built s.t. each cell M_{ij} contains the $\mathscr{P}2(i,j)$ sub-problem solution given by three components $(\alpha^*_{i,j}, \gamma^*_{i,j}, F^*_{i,j})$. The pseudo-code represents a recursive implementation of the MCKP-BFS solution. M is the matrix, N is the rows count, K_t is the columns count - 1, and k is a variable representing the number of available subchannels.

A. The UOSD Exact Solution

On the one hand, the tasks distribution sub-problem $\mathscr{P}2(i,\beta_i)$ relies on determining α_i and γ_i that correspond to the minimum cost function for SMD i. The exhaustive search over all possible solutions using a Brute Force Search that we denote (BFS-TD) is an $O(2^{n_i})$ time complexity solution. It is presented in Algorithm 2. While resolving the global problem, this solution is used in the first phase to construct matrix M. It is based on solving $N * Kt$ instances of sub-problem $\mathscr{P}2(i,\beta_i)$. Then, M is used as an input to the second sub-problem $\mathscr{P}3$.

1) Experiment 1: To investigate the feasibility and limitation of Algorithm 2, we carry the first experiment where we measure the achieved times of each phase. additionally, the construction of M is highly influenced by the distribution of the tasks' count n_i in each row, whereas the MCKP resolution is not affected. To achieve that, we vary N between 10 and 100 while we take K=600, Kt=15, and n_i in 8;9, and we use the simulation parameters described in Table II.

Algorithm 1 : MCKP-BFS

Require: M, N, K_t, and k
Ensure: the subchannels allocation $\beta^* = (\beta_1^*, \beta_2^*, ..., \beta_N^*)$;
if $k < 0$ **then**
  return $\varnothing$;
end if
if $N = 1$ **then**
  return $\beta = (min(k, K_t))$;
else
  $F^* \leftarrow \infty$;
  for $\beta_N = 0$ to K_t **do**
    $X \leftarrow$ MCKP-BFS$(M, N-1, K_t, k-\beta_N)$;
    if $X \neq \varnothing$ **then**
      $\beta^+ \leftarrow (X, \beta_N) = \left(\beta_1^+, ..., \beta_{N-1}^+, \beta_N\right)$;
      $\alpha \leftarrow \left(\alpha^*_{1,\beta_1^+}, ..., \alpha^*_{N-1,\beta_{N-1}^+}, \alpha^*_{N,\beta_N},\right)$;
      $\gamma \leftarrow \left(\gamma^*_{1,\beta_1^+}, ..., \gamma^*_{N-1,\beta_{N-1}^+}, \gamma^*_{N,\beta_N},\right)$;
      $F \leftarrow \sum_{i=1}^{N} M[i]\left[\beta_i^+\right]$;
      if $F < F^*$ **then**
        $(F^*, \beta^*) \leftarrow (F, \beta^+)$
      end if
    end if
  end for
end if

Algorithm 2 : Brute Force Search based Tasks distribution with β_i subchannel

Require: π_i, β_i
Ensure: the offloading and satisfaction vectors α_i^*, γ_i^* and the corresponding cost F^*
if $\beta_i = 0$ **then**
  $\alpha_i^* \leftarrow \boldsymbol{0}_{n_i}$;
  Build γ_i^* using α_i^*, β_i and equation (3);
  $F^* \leftarrow \mathbb{F}_i\left(\alpha_i^*, \gamma_i^*, \beta_i\right)$ according to (18);
else
  $F^* \leftarrow \infty$;
  for k=1 to $2^{n_i} - 1$ **do**
    $\alpha_i \leftarrow bin(i)$;
    Build γ_i using α_i, β_i and equation (3);
    $F \leftarrow \mathbb{F}_i\left(\alpha_i, \gamma_i, \beta_i\right)$ according to (18);
    if $F < F^*$ **then**
      $(\alpha_i^*, \gamma_i^*, F^*) \leftarrow (\alpha_i, \gamma_i, F)$
    end if
  end for
end if
return $(\alpha_i^*, \gamma_i^*, F^*)$

Fig. 3 shows the average execution time of the BFS-TD algorithm (for both values n_i = 8 and 9) and BISSA solution while we vary the total number of SMDs N between 10 and 100. In view of the obtained results, the MCKP resolution using BISSA realizes stable execution times. Indeed, for N=10 the corresponding execution times for BFS-TD(n_i=8), BFS-TD(n_i=9), BISSA are respectively 15.92, 69.87, 0.01ms. For N=50 they respectively reach 264.9,5 393.76 and 0.07ms. For N=100 they respectively reach 244.95, 383.76 and 5.61ms. It shows also an important increasing of the execution time w.r.t. N. Accordingly, this experiment shows that the exact resolutions of the first sub-problem is time consuming especially for important values of n_i and highly exceeds the MCKP resolution time using BISSA. Subsequently, an approximate solution is solicited which is the concern of the following section.

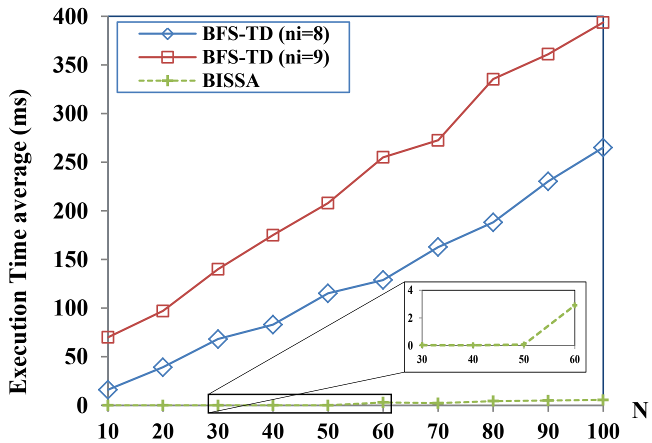


Fig. 3. Execution Time with N; K=600, Kt=15, $n_i \in \{8; 9\}$

B. The UOSD Heuristic Solution

On the other hand, given the binary form of the problem and the small decision time constraint, we propose a Simulated Annealing based heuristic solution. This heuristic optimization technique is characterized by its simplicity and general applicability features while being very efficient in terms of speed compared to other techniques. Probabilistically, this algorithm accepts not only cost gain, but also cost degradation in order to leave the local minima.

Accordingly, we propose to implement a relevant variant: Very Fast Simulated Annealing [18] which we denote VFSA-TD. Different to the classical Simulated Annealing (SA) algorithm which we adopted in our previous work [12] which denote SA-TD, This variant is characterized with a small convergence time. In this algorithm, the thermodynamic system's energy is represented by the cost function $\mathbb{F}_i$. During the solutions' space probabilistic iteration, the acceptance of the current state is done according to the following principle: we obviously accept the new state when its energy is less than its previous energy; otherwise, the new state is accepted when the probability $min\left\{exp\left(\frac{F^*-F_{new}}{T}\right),1\right\}$ is greater than a random value p. Here p is picked from a uniform distribution $U[0,1]$. Besides, a decreasing temperature, leads to a decrease in the probability for the system to shift to a new state. The temperature schedule in the SA-TD and the VFSA-TD algorithms are respectively given by:

$$T_k = T_0(a^k) \tag{23}$$

$$T_k = T_0 exp\left(-0.5k^{\frac{1}{2n_i}}\right) \tag{24}$$

Here, k is the current iteration number, a is a decreasing factor s.t. $0.5 < a < 1$. The detail of the solution is presented in Algorithm (3).

Algorithm 3 : Fast Simulated Annealing Tasks Distribution with β_i subchannel

Require: π_i,β_i,k^{max},T_0, and an initial non-empty offloading vector α_0
Ensure: the offloading and satisfaction vectors α_i^*, γ_i^* and the corresponding cost F^*

```
1:  if β_i=0 then
2:      α_i* ← 0_{n_i};
3:      Build γ_i* using α_i*,β_i and equation (3);
4:      F* ← 𝔽_i(α_i*,γ_i*,β_i) according to (18);
5:  else
6:      α_i* ← α_i ← α_0;
7:      Build γ_i using α_i,β_i and equation (3);
8:      Calculate F* = 𝔽_i(α_i*,γ_i*,β_i) according to (18);
9:      for k=1 to k^max do
10:         T ← T_0 e^{-0.5 k^{1/(2n_i)}};
11:         α_new ← rand_neighbour(α_i);
12:         if α_new ≠ 0_{n_i} then
13:             α_i ← α_new;
14:             Build γ_new using α_new,β_i and equation (3);
15:             F_new ← 𝔽_i(α_new,γ_new,β_i) according to (18)
16:             if min{ e^{(F* - F_new)/T}, 1 } ⩾ random(0,1) then
17:                 (α_i*,γ_i*,F*) ← (α_new,γ_new,F_new)
18:             end if
19:         end if
20:     end for
21: end if
22: return (α_i*,γ_i*,F*)
```

As input, Algorithm 3 requires the parameters' vector π_i of SMD i, the allocated subchannel(s) number β_i, the maximum iterations count parameter k^{max}, the initial temperature value T_0, and an initial offloading decision vector α_0. In lines 1 to 5, we handle the all-local case. We build the tasks' satisfaction vector (line 7), then we initialize the optimal cost F^*(line 8). Then we use a for loop (line 9) to repeat the annealing process using k^{max} iterations. At each step, the temperature value T is updated (line 10); then, we generate a neighboring state α_{new} of the current state α_i (line 11). If it is non-null, we build its corresponding tasks' satisfaction vector (line 14), then we calculate the new cost F_{new}(line 15). Then, we try to accept the new state using a probabilistic test(lines 16 to 18). Here, $random(0,1)$ is a function's call that uniformly generates a random number in $[0,1]$.

Despite the exponential temporal complexity of the BFS-TD method compared to the pseudo-linear complexity of SA-based methods, the BFS-TD temporal performance is acceptable for moderate values of the number of tasks n_i. Even, they are largely superior for the values in the interval [1,8]. Subsequently, to exploit this fact, we introduce an integer threshold parameter N_t. It is used s.t. if $n_i > N_t$ we use the SA based approximate heuristic solutions; otherwise, we use the exact Brute Force Search solution.

V. Evaluation and Results

In this section, we present the proposed experiments used to evaluate our proposed solutions. We were mainly based on the execution time and the cost function metrics. An algorithm's *execution time* is given by its averaged running-time, while its *cost function* is given by the averaged realizations of the cost function $\mathbb{F}$. Besides, all presented results in this work are averaged with 100 times executions.

A. Simulation Setup

All developed C++ simulation programs were built with GCC version 6.4.0. and run using a 2.4GHz Intel Core i5 processor in a PC with a maximum 8GB of RAM. Moreover, the basic parameters of the simulation experiments are listed in Table II.

TABLE II. SIMULATIONS' PARAMETERS

Parameter	values
n_i	$[2, 13]$
$D_{i,0}$	$[0.2, 2]$ MB
$\Lambda_{i,0}$	$[1, 2]$ GCycle
L_i^j	$[10, 1000]$ seconds
f_i^L	$[100, 300]$ MHz
$f_{i,k}^S$	$[3, 4]$ GHz
ξ_i^L	$[1, 10] * 10^{-10} s^{-2}$
$\xi_{i,k}^S$	$[9, 40] * 10^{-11} s^{-2}$
P_i^T	0.1 Watt
r_i	$[200, 400]$ Kb/s
$\delta_{i,k}$	$\{0\} \cup [1, 2] * 10^{-3}$ s/Kb
T_0	300

B. Experiment 2

The second experiment studies the performance of the BISSA algorithm compared to the optimal MCKP-BFS algorithm. In this experiment we vary the SMDs' number (N) between 2 and a maximum feasible experimentation value $N = 8$. Fig. 4 shows the averaged two metrics (per-SMD cost function and execution time) w.r.t. the total number of SMDs N. The right side of this figure shows the variation of the execution time of the proposed solutions. Accordingly, the BISSA solution realizes stable averaged execution times. In fact, with $n_i = 4$ and $N = 8$ it reaches only 0.42; whereas the optimal MCKP-BFS solution attains $185696.81ms$. Similarly, with $n_i = 8$ and $N = 8$ it reaches $5.37ms$; whereas the optimal MCKP-BFS solution attains $143568.47ms$. Besides, the cost function achievement for these solutions is shown in the left part of this figure. It shows that the obtained results are the same for both solutions. Indeed, this experiment shows that the more is the number of SMDs the more is the execution time high, especially with an exponential variation for the optimal MCKP-BFS solution, and a stable execution times for BISSA.

C. Experiment 3

Now, we introduce the third experiment where we study the solutions' performance related to sub-problem $\mathscr{P}2\,(i, \beta_i)$. Thus, for one SMD only, we vary n_i between 2 and 30 while we record the average of the minimum of the cost function using the optimal BFS-TD method, the SA-TD, and the VFSA-TD heuristic methods. In each case we show the averaged execution time. Fig. 5 and 6 depict the experiment's results. They show the cost function besides the execution time w.r.t. n_i.

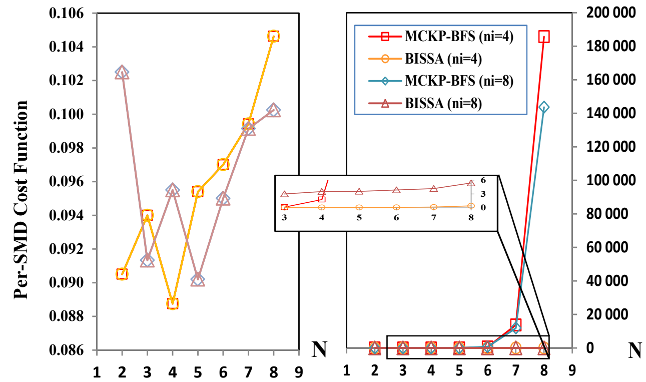


Fig. 4. Cost function and Execution Time with N; K=100, Kt=15, $n_i \in \{4; 8\}$

Hence, according to Fig. 5 that shows the obtained results in terms of Cost achievements, a small distance separating the results of all three methods is noticed.

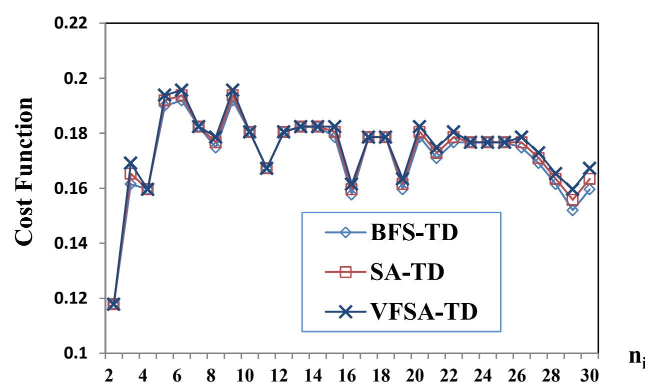


Fig. 5. Cost function performance $n_i \in [2; 30]$

Moreover, Fig. 6 shows the execution's time achievements. The obtained results illustrate important times' values for the BFS-TD method. Indeed, for $n_i = 20$ and 30 the execution time attains respectively 121.45 and $201544.20ms$. In this part of the figure and for clarity reason, we zoomed the figure to show the achievements of the heuristic solutions. as reported in the figure, for $n_i = 20$ and 30, both SA-TD and VFSA-TD heuristic methods give a stable averaged execution time that respectively attains only 0.030 and 0.086 ms for SA-TD, and 0.024 and 0.072 ms for VFSA-TD. Furthermore, as mentioned before, the BFS-TD achievements for small values of n_i between 2 and 8 in terms of execution time outrun both heuristic methods' performance. Consequently, the threshold $N_t = 8$ has to be set for a logical use of the heuristic methods.

VI. CONCLUSIONS

This paper considers a user-adaptive offloading problem with resource allocation in a multi-server Mobile Edge Com-

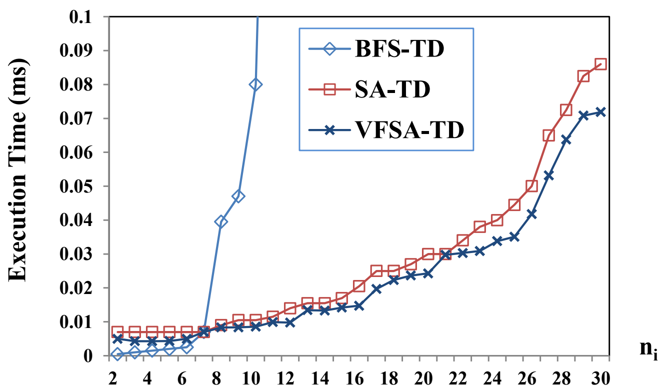


Fig. 6. Execution time performance $n_i \in [2; 30]$

puting network. We considered N smart mobile devices possessing many computation-intensive tasks each. The resulting optimization problem jointly optimizes the energy, the processing delays, and the unsatisfied processing workloads. Due to its combinatorial nature that leads to high complexity solutions, we decomposed it using a sub-problem in a first phase to build a matrix. Then, the optimal resource allocation is decided by solving a second sub-problem. To evaluate the components of the general solution, we designed a set of experiments to evaluate their performance using simulations. Accordingly, the first sub-problem exact solution is time consuming for experiments with large number of tasks. Consequently, we proposed an approximate solution based on the Very Fast Simulated Annealing algorithm. This solution is very efficient in terms of its execution time as well as its results. On the other hand, the BISSA solution for the obtained MCKP special case is much efficient and gives, in reasonable execution time, comparable result to the exact resolution. In perspectives, we plan to study services migration with backhauls' delays optimization in this proposed system.

References

[1] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[2] G. Premsankar, M. Di Francesco, and T. Taleb, "Edge computing for the internet of things: A case study," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1275–1284, 2018.

[3] M. Kadhum, S. Manaseer, and A. L. A. Dalhoum, "Cloud-edge network data processing based on user requirements using modify mapreduce algorithm and machine learning techniques," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, pp. 307–320, 2019.

[4] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 1, no. 2, pp. 89–103, 2015.

[5] X. Xu, Y. Li, T. Huang, Y. Xue, K. Peng, L. Qi, and W. Dou, "An energy-aware computation offloading method for smart edge computing in wireless metropolitan area networks," *Journal of Network and Computer Applications*, vol. 133, pp. 75–85, 2019.

[6] M.-H. Chen, B. Liang, and M. Dong, "Multi-user multi-task offloading and resource allocation in mobile cloud systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6790–6805, 2018.

[7] H. Li, "Multi-task offloading and resource allocation for energy-efficiency in mobile edge computing," *International Journal of Computer Techniques*, vol. 5, no. 1, pp. 5–13, 2018.

[8] X. Chen, "Decentralized computation offloading game for mobile cloud computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 4, pp. 974–983, 2015.

[9] K. Zhang, Y. Mao, S. Leng, Q. Zhao, L. Li, X. Peng, L. Pan, S. Maharjan, and Y. Zhang, "Energy-efficient offloading for mobile edge computing in 5g heterogeneous networks," *IEEE access*, vol. 4, pp. 5896–5907, 2016.

[10] P. Zhao, H. Tian, C. Qin, and G. Nie, "Energy-saving offloading by jointly allocating radio and computational resources for mobile edge computing," *IEEE Access*, vol. 5, pp. 11255–11268, 2017.

[11] B. Hayat, M. Nauman, S. Yousaf, M. Mehmood, and H. Saleem, "Efficient energy utilization in cloud fog environment," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 04, pp. 617–623, 2019.

[12] T. Chanyour, M. El Ghmary, Y. Hmimz, and M. O. Cherkaoui Malki, "Energy-efficient and delay-aware multitask offloading for mobile edge computing networks," *Transactions on Emerging Telecommunications Technologies*. doi: 10.1002/ett.3673.

[13] B. Gu and V. S. Sheng, "A robust regularization path algorithm for ν-support vector classification," *IEEE Transactions on neural networks and learning systems*, vol. 28, no. 5, pp. 1241–1248, 2017.

[14] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2795–2808, 2015.

[15] A. Agra and C. Requejo, "The linking set problem: a polynomial special case of the multiple-choice knapsack problem," *Journal of Mathematical Sciences*, vol. 161, no. 6, pp. 919–929, 2009.

[16] P. Sinha and A. A. Zoltners, "The multiple-choice knapsack problem," *Operations Research*, vol. 27, no. 3, pp. 503–515, 1979.

[17] E. M. Bednarczuk, J. Miroforidis, and P. Pyzel, "A multi-criteria approach to approximate solution of multiple-choice knapsack problem," *Computational Optimization and Applications*, vol. 70, no. 3, pp. 889–910, 2018.

[18] D. Pei, J. A. Quirein, B. E. Cornish, D. Quinn, and N. R. Warpinski, "Velocity calibration for microseismic monitoring: A very fast simulated annealing (vfsa) approach for joint-objective optimization," *Geophysics*, vol. 74, no. 6, pp. WCB47–WCB55, 2009.